Biased Urn Theory

Agner Fog

February 5, 2013

1 Introduction

Two different probability distributions are both known in the literature as "the" noncentral hypergeometric distribution. These two distributions will be called Fisher's and Wallenius' noncentral hypergeometric distribution, respectively.

Both distributions can be associated with the classical experiment of taking colored balls at random from an urn without replacement. If the experiment is unbiased then the result will follow the well-known hypergeometric distribution. If the balls have different size or weight or whatever so that balls of one color have a higher probability of being taken than balls of another color then the result will be a noncentral hypergeometric distribution.

The distribution depends on how the balls are taken from the urn. Wallenius' noncentral hypergeometric distribution is obtained if n balls are taken one by one. Fisher's noncentral hypergeometric distribution is obtained if balls are taken independently of each other.

Wallenius' distribution is used in models of natural selection and biased sampling. Fisher's distribution is used mainly for statistical tests in contingency tables. Both distributions are supported in the `BiasedUrn` package.

The difference between the two noncentral hypergeometric distributions is difficult to understand. I am therefore providing a detailed explanation in the following sections.

2 Definition of Wallenius' noncentral hypergeometric distribution

Assume that an urn contains N balls of c different colors and let m_i be the number of balls of color i. Balls of color i have the weight ω_i. n balls are drawn from the urn, one by one, in such a way that the probability of taking a particular ball at a particular draw is equal to this ball's fraction of the total weight of all balls that lie in the urn at this moment.

The colors of the n balls that are taken in this way will follow Wallenius' noncentral hypergeometric distribution. This distribution has the probability mass function:

$$\text{dMWNCHypergeo}(\boldsymbol{x}; \boldsymbol{m}, n, \boldsymbol{\omega}) \;=\; \left(\prod_{i=1}^{c} \binom{m_i}{x_i}\right) \int_0^1 \prod_{i=1}^{c} (1 - t^{\omega_i/d})^{x_i} \, \mathrm{d}t \;,$$

$$\text{where } d = \sum_{i=1}^{c} \omega_i (m_i - x_i) .$$

$\boldsymbol{x} = (x_1, x_2, \ldots, x_c)$ is the number of balls drawn of each color.
$\boldsymbol{m} = (m_1, m_2, \ldots, m_c)$ is the initial number of balls of each color in the urn.
$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_c)$ is the weight or odds of balls of each color.
$n = \sum_{i=1}^{c} x_i$ is the total number of balls drawn.
c is the number of colors. The unexpected integral in this formula arises as the solution to a difference equation. (The above formula is invalid in the trivial case $n = N$.)

3 Definition of Fisher's noncentral hypergeometric distribution

If the colored balls are taken from the urn in such a way that the probability of taking a particular ball of color i is proportional to its weight ω_i and the probability for each particular ball is independent of what happens to the other balls, then the number of balls taken will follow a binomial distribution for each color.

The total number of balls taken $n = \sum_{i=1}^{c} x_i$ is necessarily random and unknown prior to the experiment. After the experiment, we can determine n and calculate the distribution of colors for the given value of n. This is Fisher's noncentral hypergeometric distribution, which is defined as the distribution of independent binomial variates conditional upon their sum n.

The probability mass function of Fisher's noncentral hypergeometric distribution is given by

$$\text{dMFNCHypergeo}(\boldsymbol{x}; \boldsymbol{m}, n, \boldsymbol{\omega}) = \frac{\text{g}(\boldsymbol{x}; \boldsymbol{m}, n, \boldsymbol{\omega})}{\sum\limits_{\boldsymbol{y} \in \Xi} \text{g}(\boldsymbol{y}; \boldsymbol{m}, n, \boldsymbol{\omega})} ,$$

$$\text{where } \text{g}(\boldsymbol{x}; \boldsymbol{m}, n, \boldsymbol{\omega}) = \prod_{i=1}^{c} \binom{m_i}{x_i} \omega_i^{x_i} ,$$

$$\text{and the domain } \Xi = \left\{ \boldsymbol{x} \in \mathbb{Z}^c \,\middle|\, \sum_{i=1}^{c} x_i = n \wedge \forall\, i \in [1, c] \,:\, 0 \le x_i \le m_i \right\} .$$

4 Univariate distributions

The univariate distributions are used when the number of colors c is 2. The multivariate distributions are used when the number of colors is more than 2.

The above formulas apply to any number of colors c. The univariate distributions can be expressed by setting $c = 2,\ x_1 = x,\ x_2 = n - x,\ m_1 = m,\ m_2 = N - m,\ \omega_1 = \omega,\ \omega_2 = 1$ in the above formulas.

5 Name confusion

Wallenius' and Fisher's distribution are both known in the literature as "the" noncentral hypergeometric distribution. Fisher's distribution was first given the

name extended hypergeometric distribution, but some scientists are strongly opposed to using this name.

There is a widespread confusion in the literature because these two distributions have been given the same name and because it is not obvious that they are different. Several publications have used the wrong distribution or erroneously assumed that the two distributions were identical.

I am therefore recommending to use the prefixes Wallenius' and Fisher's to distinguish the two noncentral hypergeometric distributions. While this makes the names rather long, it has the advantage of emphasizing that there is more than one noncentral hypergeometric distribution, whereby the risk of confusion is minimized. Wallenius and Fisher are the names of the scientists who first described each of these two distributions.

The following section explains why the two distributions are different and how to decide which distribution to use in a specific situation.

6 The difference between the two distributions

Both distributions degenerate into the well-known hypergeometric distribution when all balls have the same weight. In other words: It doesn't matter how the balls are sampled if the balls are unbiased. Only if the urn experiment is biased can we get different distributions depending on how the balls are sampled.

It is important to understand how this dependence on the sampling procedure arises. In the Wallenius model, there is competition between the balls. The probability that a particular ball is taken is lower when the other balls in the urn are heavier. The probability of taking a particular ball at a particular draw is equal to its fraction of the total weight of the balls that remain in the urn at that moment. This total weight depends on the weight of the balls that have been removed in previous draws. Therefore, each draw except the first one has a probability distribution that depends on the results of the previous draws. The fact that each draw depends on the previous draws is what makes Wallenius' distribution unique and makes the calculation of it complicated. What happens to each ball depends on what has happened to other balls in the preceding draws.

In the Fisher model, there is no such dependence between draws. We may as well take all n balls at the same time. Each ball has no "knowledge" of what happens to the other balls. For the same reason, it is impossible to know the value of n before the experiment. If we tried to fix the value of n then we would have no way of preventing ball number $n+1$ from being taken without violating the principle of independence between balls. n is therefore a random variable and the Fisher distribution is a conditional distribution which can only be determined after the experiment when n is known. The unconditional distribution is c independent binomials.

The difference between Wallenius' and Fisher's distributions is low when odds ratios are near 1, and n is low compared to N. The difference between the two distributions becomes higher when odds ratios are high and n is near N.

Consider the extreme example where an urn contains one red ball with the weight 1000, and a thousand white balls each with the weight 1. We want to calculate the probability that the red ball is not taken when balls are taken one by one. The probability that the red ball is not taken in the first draw is

$\frac{1000}{2000} = \frac{1}{2}$. The probability that the red ball is not taken in the second draw, under the condition that it was not taken in the first draw, is $\frac{999}{1999} \approx \frac{1}{2}$. The probability that the red ball is not taken in the third draw, under the condition that it was not taken in the first two draws, is $\frac{998}{1998} \approx \frac{1}{2}$. Continuing in this way, we can calculate that the probability of not taking the red ball in n draws is approximately 2^{-n} for moderate values of n. In other words, the probability of not taking a very heavy ball in n draws falls almost exponentially with n in Wallenius' model. The exponential function arises because the probabilities for each draw are all multiplied together.

This is not the case in Fisher's model where balls may be taken simultaneously. Here the draws are independent and the probabilities are therefore not multiplied together. The probability of not taking the heavy red ball in Fisher's model is approximately $\frac{1}{n+1}$. The two distributions are therefore very different in this extreme case.

The following conditions must be fulfilled for Wallenius' distribution to be applicable:

- Items are taken randomly from a finite source containing different kinds of items without replacement.
- Items are drawn one by one.
- The probability of taking a particular item at a particular draw is equal to its fraction of the total weight of all items that have not yet been taken at that moment. The weight of an item depends only on its kind (color) i. (It is convenient to use the word "weight" for ω_i even if the physical property that determines the odds is something else than weight).
- The total number n of items to take is fixed and independent of which items happen to be taken.

The following conditions must be fulfilled for Fisher's distribution to be applicable:

- Items are taken randomly from a finite source containing different kinds of items without replacement.
- Items are taken independently of each other. Whether one item is taken is independent of whether another item is taken. Whether one item is taken before, after, or simultaneously with another item is irrelevant.
- The probability of taking a particular item is proportional to its weight. The weight of an item depends only on its kind (color) i.
- The total number n of items that will be taken is not known before the experiment.
- n is determined after the experiment and the conditional distribution for n known is desired.

7 Examples

The following examples will further clarify which distribution to use in different situations.

7.1 Example 1

You are catching fish in a small lake that contains a limited number of fish. There are different kinds of fish with different weights. The probability of catching a particular fish is proportional to its weight when you only catch one fish.

You are catching the fish one by one with a fishing rod. You have been ordered to catch n fish. You are determined to catch exactly n fish regardless of how long time it may take. You are stopping after you have caught n fish even if you can see more fish that are tempting you.

This scenario will give a distribution of the types of fish caught that is equal to Wallenius' noncentral hypergeometric distribution.

7.2 Example 2

You are catching fish as in example 1, but you are using a big net. You are setting up the net one day and coming back the next day to remove the net. You count how many fish you have caught and then you go home regardless of how many fish you have caught.

Each fish has a probability of getting into the net that is proportional to its weight but independent of what happens to the other fish.

This scenario gives Fisher's noncentral hypergeometric distribution after n is known.

7.3 Example 3

You are catching fish with a small net. It is possible that more than one fish can go into the net at the same time. You are using the net multiple times until you have at least n fish.

This scenario gives a distribution that lies between Wallenius' and Fisher's distributions. The total number of fish caught can vary if you are getting too many fish in the last catch. You may put the excess fish back into the lake, but this still doesn't give Wallenius' distribution. This is because you are catching multiple fish at the same time. The condition that each catch depends on all previous catches does not hold for fish that are caught simultaneously or in the same operation.

The resulting distribution will be close to Wallenius' distribution if there are only few fish in the net in each catch and you are catching many times.

The resulting distribution will be close to Fisher's distribution if there are many fish in the net in each catch and you are catching few times.

7.4 Example 4

You are catching fish with a big net. Fish are swimming into the net randomly in a situation that resembles a Poisson process. You are watching the net all the time and take up the net as soon as you have caught exactly n fish.

The resulting distribution will be close to Fisher's distribution because the fish swim into the net independently of each other. But the fates of the fish are not totally independent because a particular fish can be saved from getting caught if n other fish happen to get into the net before the time that this particular fish would have been caught. This is more likely to happen if the other fish are heavy than if they are light.

7.5 Example 5

You are catching fish one by one with a fishing rod as in example 1. You need a particular amount of fish in order to feed your family. You are stopping when the total weight of the fish you have caught exceeds a predetermined limit.

The resulting distribution will be close to Wallenius' distribution, but not exactly because the decision to stop depends on the weight of the fish you have caught so far. n is therefore not known exactly before the fishing trip.

7.6 Conclusion

These examples show that the distribution of the types of fish you catch depends on the way they are caught. Many situations will give a distribution that lies somewhere between Wallenius' and Fisher's noncentral hypergeometric distributions.

An interesting consequence of the difference between these two distributions is that you will get more of the heavy fish, on average, if you catch n fish one by one than if you catch all n at the same time.

These conclusions can of course be applied to biased sampling of other items than fish.

8 Applications

The biased urn models can be applied to many different situations where items are sampled with bias and without replacement.

8.1 `Calculating probabilities etc.`

Probabilities, mean and variance can be calculated with the appropriate functions. More complicated systems, such as the natural selection of animals, can be treated with Monte Carlo simulation, using the random variate generating functions.

8.2 `Measuring odds ratios`

The odds of a sampling process can be measured by an experiment or a series of experiments where the number of items sampled of each kind (color) is counted.

It is recommended to use sampling with replacement if possible. Sampling with replacement makes it possible to use the binomial distribution, whereby the calculation of the odds becomes simpler and more accurate. If sampling with replacement is not possible, then the procedure of sampling without replacement must be carefully controlled in order to get a pure Wallenius' distribution or a pure Fisher's distribution rather than a mixture of the two, as explained in

the examples above. Use the `odds` functions to calculate the odds ratios from experimental values of the mean.

8.3 `Estimating the number of items of a particular kind from experimental sampling`

It is possible to estimate the number of items of a particular kind, for example defective items in a production, from biased sampling. The traditional procedure is to use unbiased sampling. But a model of biased sampling may be used if bias is unavoidable or if bias is desired in order to increase the probability of detecting e.g. defective items.

It is recommended to use sampling with replacement if possible. Sampling with replacement makes it possible to use the binomial distribution, whereby the calculation of the number of items becomes simpler and more accurate. If sampling with replacement is not possible, then the procedure of sampling without replacement must be carefully controlled in order to get a pure Wallenius' distribution or a pure Fisher's distribution rather than a mixture of the two, as explained in the examples above. The value of the bias (odds ratio) must be determined before the numbers can be calculated.

Use the functions with names beginning with "`num`" to calculate the number of items of each kind from the result of a sampling experiment with known odds ratios.

9 Demos

The following demos are included in the `BiasedUrn` package:

9.1 `CompareHypergeo`

This demo shows the difference between the hypergeometric distribution and the two noncentral hypergeometric distributions by plotting the probability mass functions.

9.2 `ApproxHypergeo`

This demo shows shows that the two noncentral hypergeometric distributions are approximately equal when the parameters are adjusted so that they have the same mean rather than the same odds.

9.3 `OddsPrecision`

Calculates the precision of the `oddsWNCHypergeo` and `oddsFNCHypergeo` functions that are used for estimating the odds from a measured mean.

9.4 `SampleWallenius`

Makes 100,000 random samples from Wallenius noncentral hypergeometric distribution and compares the measured mean with the theoretical mean.

9.5 `UrnTheory`

Displays this document.

10 Calculation methods

The `BiasedUrn` package can calculate the univariate and multivariate Wallenius' and Fisher's noncentral hypergeometric distributions. Several different calculation methods are used, depending on the parameters.

The calculation methods and sampling methods are documented at `http://www.agner.org/random/theory/`.

11 References

Fog, A. (2008a). Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution. *Communications in Statistics, Simulation and Computation.* Vol. 37, no. 2, pp 258-273.

Fog, A. (2008b). Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions. *Communications in Statistics, Simulation and Computation.* Vol. 37, no. 2, pp 241-257.

Johnson, N. L., Kemp, A. W. Kotz, S. (2005). *Univariate Discrete Distributions.* Hoboken, New Jersey: Wiley and Sons.

McCullagh, P., Nelder, J. A. (1983). *Generalized Linear Models.* London: Chapman & Hall.

`http://www.agner.org/random/theory/`.