Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions

Agner Fog, 2006-06-16.

A revised version of this article will appear in *Communications in Statistics, Simulation and Computation*, vol. 37, no. 2, 2008.

SUMMARY. Several methods for generating random variables with univariate and multivariate Wallenius' and Fisher's noncentral hypergeometric distributions are developed. Methods for the univariate distributions include: simulation of urn experiments, inversion by binary search, inversion by chop-down search from the mode, ratio-of-uniforms rejection method, and rejection by sampling in the τ domain. Methods for the multivariate distributions include: simulation of urn experiments, conditional method, Gibbs sampling, and Metropolis-Hastings sampling. These methods are useful for Monte Carlo simulation of models of biased sampling and models of evolution and for calculating moments and quantiles of the distributions.

KEY WORDS: Noncentral hypergeometric distribution; Wallenius; Fisher; Multivariate distribution; Variate generation; Sampling; Simulation.

1. Introduction

Two different probability distributions are both known in the literature as "the" noncentral hypergeometric distribution. These two distributions will be called Wallenius' and Fisher's noncentral hypergeometric distribution, respectively. An accompanying paper describes the nomenclature problems as well as several methods for calculating probabilities from Wallenius' noncentral hypergeometric distribution (Fog, 2007). Wallenius' distribution has many applications, including models of biased sampling and competitive models of Darwinian evolution (Wallenius, 1963; Manly, 1974). Fisher's noncentral hypergeometric distribution may be used for modeling non-competitive models of evolution when the total number of survivors is known, as well as for statistical tests on contingency tables (McCullagh and Nelder, 1983). Methods for sampling from both distributions are needed for Monte Carlo simulations of evolutionary systems and models of biased sampling as well as for finding moments, quantiles, etc. The purpose of the present article is to develop efficient methods for sampling from these distributions.

When comparing the efficiency of different sampling methods, we want to distinguish between the *set-up time*, which is the time required to compute constants that depend only on the parameters of the distribution, and the *sampling time*, which is the time required to generate one variate, not including the set-up time (Stadlober, 1989). A simulation of an evolutionary system will typically require only one variate for each set of parameters, because the composition of the gene pool is likely to change for each generation. The best sampling method for this application will thus optimize the total execution time, i.e. the sum of the set-up time and the sampling time. Other applications that require many variates for each set of parameters will prefer a method that optimizes the sampling time only, while the set-up time is less important.

An implementation of the methods developed in this article in the C++ programming language is available from www.agner.org/random.

2. Definition of distributions

The multivariate Wallenius' noncentral hypergeometric distribution has the probability function given by (Fog, 2007; Chesson, 1976):

$$\mathrm{mwnchypg}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) = \Lambda(\mathbf{x})\mathrm{I}(\mathbf{x}), \text{ where} \tag{1}$$

$$\Lambda(\mathbf{x}) = \prod_{i=1}^{c} \binom{m_i}{x_i}, \tag{2}$$

$$\mathrm{I}(\mathbf{x}) = \int_0^1 \prod_{i=1}^{c} (1 - t^{\omega_i / d})^{x_i} \, dt, \tag{3}$$

$$d = \boldsymbol{\omega} \cdot (\mathbf{m} - \mathbf{x}) = \sum_{i=1}^{c} \omega_i (m_i - x_i), \tag{4}$$

$$\mathbf{x} = (x_1, x_2, ..., x_c), \quad \mathbf{m} = (m_1, m_2, ..., m_c), \; \boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_c), \tag{5}$$

which is valid for $d > 0$.

The univariate distribution ($c = 2$) is defined as

$$\mathrm{wnchypg}(x; n, m, N, \omega) = \mathrm{mwnchypg}((x_1, x_2); n, (m_1, m_2), (\omega_1, \omega_2)), \tag{6}$$

where $x_1 = x$, $x_2 = n\text{-}x$, $m_1 = m$, $m_2 = N\text{-}m$, $\omega_1 = \omega$, $\omega_2 = 1$.

The multivariate Fisher's noncentral hypergeometric distribution, which is also called the extended hypergeometric distribution, is defined as the conditional distribution of independent binomial variates given their sum (Harkness, 1965). The probability function is (McCullagh and Nelder, 1983):

$$\mathrm{mfnchypg}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) = \frac{\mathrm{g}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega})}{\sum_{\mathbf{y} \in \mathcal{S}} \mathrm{g}(\mathbf{y}; n, \mathbf{m}, \boldsymbol{\omega})}, \text{ where} \tag{7}$$

$$\mathrm{g}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) = \prod_{i=1}^{c} \binom{m_i}{x_i} \omega_i^{x_i}, \text{ and the support} \tag{8}$$

$$\mathcal{S} = \left\{ \mathbf{x} \in \mathbb{Z}_{0+}^{c} \mid \sum_{i=1}^{c} x_i = n \right\}. \tag{9}$$

The univariate distribution ($c = 2$) has the probability function

$$\mathrm{fnchypg}(x; n, m, N, \omega) = \mathrm{mfnchypg}\{(x_1, x_2); n, (m_1, m_2), (\omega_1, \omega_2)\}, \tag{10}$$

where $x_1 = x$, $x_2 = n\text{-}x$, $m_1 = m$, $m_2 = N\text{-}m$, $\omega_1 = \omega$, $\omega_2 = 1$.

The parameterization here is chosen so as to emphasize the similarity between the two distributions. Both distributions are reduced to the (multivariate) binomial distribution when $n = 1$, or to the (multivariate) hypergeometric distribution when all ω_i's are equal. Hence, it is not surprising that the two distributions approximate each other when $n \ll N$ and when the odds ratios are all close to 1. The univariate Fisher's distribution has the same minimum and maximum as the (central) hypergeometric distribution:

$$x_{\min} = \max(0, n + m - N), \quad x_{\max} = \min(n, m). \tag{11}$$

The following symmetry relations are easily derived:

$$\text{fnchypg}(x;n,m,N,\omega) = \text{fnchypg}(n-x;n,N-m,N,1/\omega)\,, \tag{12}$$

$$\text{fnchypg}(x;n,m,N,\omega) = \text{fnchypg}(x;m,n,N,\omega)\,, \tag{13}$$

$$\text{fnchypg}(x;n,m,N,\omega) = \text{fnchypg}(m-x;N-n,m,N,1/\omega)\,. \tag{14}$$

(11), and (12) apply analogously to the univariate Wallenius' distribution; (13) and (14) do not.

In the following, the same abbreviations are used for the distributions and their probability functions.

3. Sampling from the univariate Fisher's noncentral hypergeometric distribution

When sampling from the univariate Fisher's noncentral hypergeometric distribution, it may be advantageous to apply the symmetry transformations (12) (13) (14), if needed, to make $n \le m \le N/2$ so that $x_{\min} = 0$ and $x_{\max} = n$.

3.1 Direct inversion method and chop-down search from the mode

Let U be a variate with uniform distribution on [0,1), let f(x) be the probability function of a discrete distribution, and let $\mathrm{F}(x) = \sum_{y=x_{\min}}^{x} \mathrm{f}(y)$ be the cumulative distribution function. The smallest x that satisfies $U < \mathrm{F}(x)$ will then have the distribution f (Devroye, 1986; Cheng, 1998).

A method based on binary search in a table of F(x) is advantageous in applications where sampling time is more important than set-up time. When set-up time is important, it is preferable to minimize the number of f(x) values that have to be calculated by evaluating the most probable x-values first. Let M denote the mode, and define the mapping $z_0 = M,\ z_1 = M+1,\ z_2 = M-1,\ z_3 = M+2,\ \ldots$, where z-values beyond $x_{\min}$ and $x_{\max}$ are excluded. If Y is the smallest number that satisfies $U < \sum_{j=0}^{Y} \mathrm{f}(z_j)$, then z_Y will have the distribution f (Stadlober, 1989; Devroye, 1986). Liao and Rosen (2001) use this method for sampling from Fisher's noncentral hypergeometric distribution, using the recurrence relation

$$\text{fnchypg}(x;n,m,N,\omega) = \text{fnchypg}(x-1;n,m,N,\omega)\frac{(m-x+1)(n-x+1)}{x(N-m-n+x)}\omega\,. \tag{15}$$

This method still has a large set-up time because the g function (8) has to be evaluated for all x-values in the support. Liao and Rosen (2001) recommend to calculate the g values relative to g(M) using (15) and ignore negligible values in the tails. Noting that division takes 5 - 10 times as long as multiplication on contemporary computers, we may improve the speed of this method further by scaling the g values and their sum with the product of all denominators in (15) in order to avoid the many divisions. During the course of these calculations, it may be necessary to downscale all values to avoid numerical overflow.

3.2 Ratio of uniforms rejection method

The principle of the rejection method is as follows. Let f(x) and h(x) be two probability functions with the same domain Ξ so that

$$\exists k \in \mathbb{R}_+ \forall x \in \Xi : \mathrm{f}(x) \le k\,\mathrm{h}(x)\,. \tag{16}$$

Generate a variate X with distribution h and a Bernoulli variate Z with probability parameter p(X) = f(X) / {kh(X)}. Repeat this procedure until Z = 1. Then X will have the distribution f (Devroye, 1986). The ratio-

of-uniforms method is a rejection method based on the hat function (Stadlober, 1990):

$$\mathrm{h}(x) = \begin{Bmatrix} 1 & \text{for } a - s \le x \le a + s \\ \dfrac{s^2}{(x-a)^2} & \text{elsewhere} \end{Bmatrix} \tag{17}$$

Let U and V be two independent random numbers with uniform distribution in the intervals $0 < U \le 1$ and $-1 \le V \le 1$, respectively. Do the transformation $X = sV/U + a$, $Y = U^2$. The rectangle in the (u,v) plane is thereby transformed into the area delimited by the hat function $y = \mathrm{h}(x)$ in the (x,y) plane. The acceptance condition for a discrete distribution is $kV < \mathrm{f}(\lfloor X \rfloor)$. a, s and k are chosen so that $\mathrm{f}(x) \le k\,\mathrm{h}(x)$ for all $x \in \Xi$. The advantage of the ratio-of-uniforms method is that the generation of X and Y is simple and fast, and the acceptance rate is reasonably good. There is no reason to normalize $\mathrm{f}(x)$ and $\mathrm{h}(x)$ because any proportionality factor is absorbed by the optimal choice of $k = \mathrm{f}(M)$, where M is the mode. This is a tremendous advantage when sampling from the fnchypg distribution because we do not have to calculate the large sum in the denominator of (7). We may improve the speed further by removing all factors that do not depend on x and replace $\mathrm{f}(x)$ by

$$\mathrm{g}_1(x) = \frac{\mathrm{g}(x)}{m_1!\,m_2!} = \prod_{i=1}^{2} \frac{\omega_i^{\,x_i}}{x_i!\,(m_i - x_i)!}\,. \tag{18}$$

The optimal value of a is $\mu + \frac{1}{2}$ for a symmetric mass function. It is fairly straightforward to find the optimal value of s by numerical methods for a given set of parameters (Stadlober, 1989). However, since many applications require higher priority to minimizing the set-up time than to improving the acceptance rate, we will prefer a sub-optimal value of s that can be calculated relatively fast. Ahrens and Dieter (1989) find that the following values of a and s fit the Poisson distribution

$$a = \mu + \tfrac{1}{2}, \quad s = \sqrt{\frac{2}{e}\left(\sigma^2 + \tfrac{1}{2}\right)} + \frac{3}{2} - \sqrt{\frac{3}{e}}\,, \tag{19}$$

where μ and σ are the mean and standard deviation. Stadlober (1989, 1990) has found experimentally (but without theoretical proof) that the same formula fits the binomial and hypergeometric distributions. It is found experimentally that this formula also fits Fisher's noncentral hypergeometric distribution when the exact mean and variance are used. For practical reasons, we prefer to use the following approximations to the mean and variance (Levin, 1984; Liao, 1992):

$$(m-\mu)(n-\mu)\omega \;\approx\; \mu(N-m-n+\mu), \tag{20}$$

$$\sigma^2 \;\approx\; \frac{N}{N-1} \Big/ \left(\frac{1}{\mu} + \frac{1}{m-\mu} + \frac{1}{n-\mu} + \frac{1}{\mu + N - m - n} \right). \tag{21}$$

When these approximations are used, the formula needs to be modified to

$$s = h_1 + h_2\sqrt{\sigma^2 + \tfrac{1}{2}} + h_3\left|\log \omega\right| \quad \text{with } h_1 = 0.514,\ h_2 = 0.8585,\ h_3 = 0.016. \tag{22}$$

In general, a dominating function $\mathrm{h}(x)$ may be obtained either by theoretical or by experimental means. Universal theoretical methods either have long set-up times or poor acceptance rates (Devroye, 1986). While it is possible that an experimentally obtained dominating function gives a tighter fit than an expression obtained by theoretical methods, we have to be strict about the criteria for accepting an experimental verification. Stadlober does not specify any such criteria. The number of possible parameter sets to test is infinite if no upper bound for the integer-valued parameters is specified, or if at least one

parameter is real-valued, as is the case here. A natural approach is therefore to test the validity of the formula for a large number of random parameter sets. The distribution from which the random parameter sets are sampled may be critical. It is possible that a hypothetical parameter set that invalidates the formula has a very low probability for a given distribution of parameter sets. It is therefore necessary to repeat the test procedure with different distributions. Obvious choices include combinations of uniform and exponential distributions for each parameter. Furthermore, it is recommended to study the system in order to identify any narrow areas of the parameter space that are particularly critical. If such critical areas are found, then it is necessary to design a distribution that increases the probability that parameter sets fall in these areas. This has been the case with Wallenius' distribution, as discussed below.

The formula (22) has been verified by testing with several different distributions totalling more than 10^9 random parameter sets in the range $N \le 10^9 \wedge 10^{-9} \le \omega \le 10^9$.

The calculation of the mean according to (20) is subject to loss of precision when n, m and ω are all very large. It is therefore recommended to use the symmetry transformations (12) and (14), if needed, to make $n \le N/2$ and $m \le N/2$.

4. Sampling from the multivariate Fisher's noncentral hypergeometric distribution

In special cases, the number of colors can be reduced by eliminating colors with zero weight or zero number, or by joining colors with the same weight:

$$\begin{aligned}
&\mathrm{mfnchypg}\{(x_1,...,x_{c-1},0);n,(m_1,...,m_{c-1},0),(\omega_1,...,\omega_{c-1},\omega_c)\} = \\
&\mathrm{mfnchypg}\{(x_1,...,x_{c-1},0);n,(m_1,...,m_{c-1},m_c),(\omega_1,...,\omega_{c-1},0)\} = \\
&\mathrm{mfnchypg}\{(x_1,...,x_{c-1});n,(m_1,...,m_{c-1}),(\omega_1,...,\omega_{c-1})\}
\end{aligned} \tag{23}$$

$$\begin{aligned}
&\mathrm{mfnchypg}\{(x_1,...,x_{c-1},x_c);n,(m_1,...,m_{c-1},m_c),(\omega_1,...,\omega_{c-1},\omega_{c-1})\} = \\
&\mathrm{mfnchypg}\{(x_1,...,x_{c-1}+x_c);n,(m_1,...,m_{c-1}+m_c),(\omega_1,...,\omega_{c-1})\}\mathrm{hypg}(x_c;x_{c-1}+x_c,m_c,m_{c-1}+m_c)
\end{aligned} \tag{24}$$

4.1 Conditional method

The conditional method for sampling from a multivariate distribution is the method where X_1 is sampled first from the marginal distribution, then X_2 is sampled according to the conditional distribution of X_2 given X_1, and so forth (Johnson, 1987). This method is useful for the multivariate (central) hypergeometric distribution, where the marginal distribution of X_1, as well as the conditional distribution of X_2, etc., are univariate hypergeometric distributions. For Fisher's noncentral hypergeometric distribution, however, the marginal distribution of X_1 is difficult to calculate exactly, so we have to do with approximations.

Consider, for the sake of argument, the four-color example where $\omega_1 = \omega_2 \ne \omega_3 = \omega_4$. Here, the distribution of X_1+X_2 is a univariate fnchypg distribution according to (24), while the marginal distribution of X_1 alone is more difficult to calculate. Therefore, in the case where $\omega_1 \approx \omega_2 \gg \omega_3 \approx \omega_4$, we expect a method based on an approximation to the marginal distribution of X_1+X_2 to be more accurate than a method based on an approximation to the marginal distribution of X_1 alone.

Assume that any possibility for reducing the number of colors according to (23) and (24) has been utilized, and let the colors be sorted by weight so that $\omega_1 > \omega_2 > ... > \omega_c$. Define the geometric mean of the highest and the lowest weight

$$\omega^{\blacktriangle} = \sqrt{\omega_1 \omega_c} \,. \tag{25}$$

Let b be the lightest color not lighter than $\omega^{\blacktriangle}$:

$$\omega_b \geq \omega^{\blacktriangle} > \omega_{b+1}. \tag{26}$$

Consider colors not lighter than $\omega^{\blacktriangle}$ as the heavy group H = {1,2,...,b}, and colors lighter than $\omega^{\blacktriangle}$ as the light group L = {b+1,b+2,...,c}. The number of balls in the sample that belong to the heavy color group and the light color group, respectively, are

$$Y_{\mathrm{H}} = \sum_{i=1}^{b} X_i \quad \text{and} \quad Y_{\mathrm{L}} = \sum_{i=b+1}^{c} X_i = n - Y_{\mathrm{H}}. \tag{27}$$

The distribution of Y_{H} is given by the probability function

$$\mathrm{f}_{\mathrm{H}}(y) = \Pr(Y_{\mathrm{H}} = y) = \sum_{\mathbf{x} \in \Xi_{by}} \mathrm{f}(\mathbf{x}) \quad \text{where} \tag{28}$$

$$\Xi_{by} = \left\{ \mathbf{x} \in \Xi \mid \sum_{i=1}^{b} x_i = y \right\}. \tag{29}$$

The mean weight of each group is

$$\omega_{\mathrm{H}} = \frac{\sum_{i=1}^{b} m_i \omega_i}{m_{\mathrm{H}}}, \quad \omega_{\mathrm{L}} = \frac{\sum_{i=b+1}^{c} m_i \omega_i}{m_{\mathrm{L}}}, \text{ where} \tag{30}$$

$$m_{\mathrm{H}} = \sum_{i=1}^{b} m_i, \quad m_{\mathrm{L}} = \sum_{i=b+1}^{c} m_i. \tag{31}$$

The distribution of Y_{H} can now be approximated by

$$\mathrm{f}_{\mathrm{H}}(y) \approx \mathrm{fnchyp}(y; n, m_{\mathrm{H}}, N, \omega_{\mathrm{H}} / \omega_{\mathrm{L}}), \quad N = \sum_{i=1}^{c} m_i = m_H + m_L. \tag{32}$$

Define the subvectors

$$\begin{aligned} &\mathbf{X}_{\mathrm{H}} = (X_1,..., X_b), \ \mathbf{X}_{\mathrm{L}} = (X_{b+1},..., X_c), \\ &\mathbf{m}_{\mathrm{H}} = (m_1,..., m_b), \ \mathbf{m}_{\mathrm{L}} = (m_{b+1},..., m_c), \\ &\boldsymbol{\omega}_{\mathrm{H}} = (\omega_1,..., \omega_b), \ \boldsymbol{\omega}_{\mathrm{L}} = (\omega_{b+1},..., \omega_c). \end{aligned} \tag{33}$$

Now it follows from the definition of Fisher's noncentral hypergeometric distribution that the conditional distribution within each group, given $Y_{\mathrm{H}} = y$, are

$$\mathbf{X}_{\mathrm{H}} \sim \mathrm{mfnchyp}(y, \mathbf{m}_{\mathrm{H}}, \boldsymbol{\omega}_{\mathrm{H}}) \quad \text{and} \tag{34}$$

$$\mathbf{X}_{\mathrm{L}} \sim \mathrm{mfnchyp}(n - y, \mathbf{m}_{\mathrm{L}}, \boldsymbol{\omega}_{\mathrm{L}}). \tag{35}$$

We can now obtain a sample that approximates the multivariate distribution by sampling Y_{H} from the univariate distribution (32) and apply the same procedure recursively to the distributions within subgroups according to (34) and (35), until all X_i's have been determined. It follows from the above arguments that the distribution of the resulting $\mathbf{X}$ approaches the exact distribution $\mathrm{mfnchyp}(n, \mathbf{m}, \boldsymbol{\omega})$ as the differences in weight within groups go towards zero. This has been confirmed experimentally, where a reasonable precision was obtained when differences in weight within groups was not too large.

4.2 Gibbs sampling

A Gibbs sampler is an infinite Markov Chain sequence whose limit is a random vector with the desired distribution (Casella and George, 1992). Suppose we want to generate a random vector $\mathbf{X} = (X_1,...,X_c)$ with distribution f. The sequence $\{ {}^{k}\mathbf{X},\ k = 0...\infty \}$ is defined by an arbitrary starting point ${}^{0}\mathbf{X}$ and a transition from ${}^{k}\mathbf{X}$ to ${}^{k+1}\mathbf{X}$ which is created as follows. Obtain ${}^{k+1}X_1$ from the conditional distribution of X_1 given $X_2, X_3, ... X_c$. Then obtain ${}^{k+1}X_2$ from the conditional distribution of X_2 given $X_1, X_3, X_4,... X_c$, and so forth. In short, for $i = 1...c$, obtain ${}^{k+1}X_i$ by drawing from the conditional distribution

$$ {}^{k+1}X_i \sim \mathrm{f}(x_i \mid \forall j < i : x_j = {}^{k+1}x_j, \forall j > i : x_j = {}^{k}x_j) . \qquad (36) $$

This makes it possible to obtain variates of a multivariate distribution by sampling from the c conditional distributions of each component given the other components. The transition from ${}^{k}\mathbf{X}$ to ${}^{k+1}\mathbf{X}$ by repetition of (36) for $i = 1...c$ is called a scan. In order to achieve independence of the starting point ${}^{0}\mathbf{X}$ we need a burn-in period of many scans before accepting a random vector. A Gibbs sampler for a discrete distribution is convergent if all states communicate, and the limiting distribution is exact for $k \to \infty$ (Roberts and Polson 1994; Besag et. al. 1995). McDonald, Smith and Forster (1999) have mentioned the possibility of using a Gibbs sampler for the multivariate Fisher's noncentral hypergeometric distribution, but lacking an efficient way of sampling from the univariate distribution, they apparently have not implemented it. Gibbs samplers have been developed mainly for calculating moments and quantiles of non-standard distributions. The interdependence of consecutive ${}^{k}\mathbf{X}$ is not a problem in such applications. For simulation applications, however, all samples must be independent. We therefore need a new starting point and a new burn-in period for each sample. It is therefore important to obtain fast convergence in order to keep the burn-in period short.

A multivariate Fisher's noncentral hypergeometric distribution with c colors can be regarded as a c-1 dimensional distribution because for any color j we can calculate $X_j = n - \sum_{i \neq j} X_i$. The conditional distribution of X_i and X_j given the remaining components is a univariate Fisher's noncentral hypergeometric distribution, according to (34), hence

$$ {}^{k+1}X_i \sim \mathrm{fnchyp}({}^{k}X_i + {}^{k}X_j, m_i, m_i + m_j, \omega_i / \omega_j), \quad {}^{k+1}X_j = {}^{k}X_i + {}^{k}X_j - {}^{k+1}X_i . \qquad (37) $$

j may be fixed or variable. We can improve the convergence by sampling the colors by the order of their variance. The variance of the marginal distribution of X_i can be approximated by the variance of a univariate distribution with the same mean, using (21):

$$ \sigma_i^2 \approx \frac{N}{N-1} \Bigg/ \left(\frac{1}{\mu_i} + \frac{1}{m_i - \mu_i} + \frac{1}{n_i - \mu_i} + \frac{1}{\mu_i + N - m_i - n} \right) . \qquad (38) $$

The mean of a multivariate Fisher's noncentral hypergeometric distribution can be approximated by the multivariate extension to (20):

$$ \mu_i = \frac{m_i r \omega_i}{r \omega_i + 1} \text{ where } r \text{ is the unique positive solution to } \sum_{i=1}^{c} \mu_i = n . \qquad (39) $$

Other expressions for the mean and variance are given by McCullagh and Nelder (1983).

In order to minimize the burn-in period, it is useful to use the approximation obtained by the conditional method as starting point, and make each scan through consecutive i with $j = i \bmod c$. It is necessary that m_i and ω_i are positive for all i in order to make sure that all possible states communicate. The fact that this algorithm will converge towards the exact distribution follows from the theory of Metropolis-Hastings sampling (Hastings, 1970). Alternatively, we may define the mapping

$$Y_1 = X_1 + X_2,\ Y_2 = X_2 + X_3, ..., Y_c = X_c + X_1 \tag{40}$$

and apply Gibbs sampler theory (Roberts and Polson, 1994) to the distribution of $(Y_1,...,Y_c)$, but (40) is injective only for *c* odd.

5. Sampling from Wallenius' noncentral hypergeometric distribution

5.1 Simulating the urn experiment

An obvious method for generating variates with Wallenius' noncentral hypergeometric distribution is to simulate an urn experiment with bias and without replacement. This method is useful for both the univariate and the multivariate distribution.

It is required that the balls be taken one by one. The multivariate Wallenius' noncentral hypergeometric distribution is reduced to the multinomial distribution for $n = 1$:

$$\text{mwnchyp}(\mathbf{x};1,\mathbf{m},\boldsymbol{\omega}) = \text{p}(i) = \frac{m_i\omega_i}{\sum_{j=1}^{c} m_j\omega_j} \quad \text{for} \quad x_i = 1 \wedge \forall j \neq i : x_j = 0. \tag{41}$$

The probability distribution of the color of the first ball is p(*i*). A variate *I* with this distribution is generated by inversion, using a table of the cumulative distribution function

$$\text{F}(i) = \sum_{j=1}^{i} \text{p}(j). \tag{42}$$

The number of balls of each color that remain in the urn after the first ball has been taken is found by decrementing m_I. The distribution of the second ball is found in the same way, using the adjusted value of **m**. This process is repeated until *n* balls have been taken. This method requires up to *n* uniform variates. It is therefore economical only for small *n*. In cases where *c* is very high, it is advantageous to organize the F(*i*) table as a binary tree in order to minimize the number of p(*i*) values that have to be recalculated after each draw.

5.2 Direct inversion method and chop-down search from the mode

The inversion methods described above are also applicable to the univariate Wallenius noncentral hypergeometric distribution. The standard deviation σ of Wallenius' noncentral hypergeometric distribution is often quite small. As a rule of thumb, we may say that the variance of the discrete probability distributions decrease in the order: Poisson > binomial > hypergeometric > Fisher's noncentral hypergeometric > Wallenius' noncentral hypergeometric. A small σ means that relatively few probability values need to be calculated when searching from the mode. On the other hand, the calculation time for the probability function is quite high. Inversion is therefore economical only when σ is low or when many variates are required with the same set of parameters.

5.3 Ratio of uniforms rejection method

While Ahrens and Dieter's formula (19) for the ratio-of-uniforms rejection method fits Fisher's noncentral hypergeometric distribution, it does not fit Wallenius' distribution. The following corrections are required in order to make sure that $\text{f}(x) \le k\ \text{h}(x)$ for Wallenius' distribution:

$$a = \mu^* + \tfrac{1}{2} \tag{43}$$

$s = s_1 + s_2 + s_3 + s_4$, where (44)

$$s_1 = h_1,\ s_2 = h_2\sqrt{\sigma_N^2 + \tfrac{1}{2}},\ s_3 = h_3\left|M - \mu^*\right|,\ s_4 = \frac{h_4 N^{h_5}}{\max(1,\alpha)^2}, \tag{45}$$

$$\alpha = \min(x_{\max} - \mu^* - s_1 - s_2 - s_3, \mu^* - s_1 - s_2 - s_3 - x_{\min}), \text{ and} \tag{46}$$

$$h_1 = 0.40,\ h_2 = 0.8579,\ h_3 = 0.40,\ h_4 = 0.029,\ h_5 = 0.23, \tag{47}$$

where M is the mode, μ^* is the approximation to the mean given by the equation (Fog, 2007)

$$\left(1 - \frac{n-\mu^*}{N-m}\right)^{\omega} = 1 - \frac{\mu^*}{m}, \tag{48}$$

and σ_N is the approximation to the standard deviation given by

$$\sigma_N = \frac{1}{\mathrm{f}(M)\sqrt{2\pi}}. \tag{49}$$

It is not a problem here that no good way of approximating the standard deviation is known, because the necessary corrections to Ahrens and Dieter's formula are smaller, and the acceptance rate better, when the crude approximation σ_N is used than when the exact standard deviation is used. s_3 and s_4 are only needed in situations where the mode is near $x_{\min}$ or $x_{\max}$. s_4 is not needed when $\frac{1}{5} < \omega < 5 \vee \alpha < -\frac{1}{2} \vee \alpha > 8$. As mentioned above in connection with Fisher's noncentral hypergeometric distribution, it is necessary to test the experimentally determined dominating function thoroughly. This is particularly important for Wallenius' distribution because the most critical parameter sets are concentrated in relatively small areas of the parameter space characterized by n being close to N. The formula (43)-(47) has therefore been verified by testing with several different distributions of random parameter sets including distributions that give a high probability of n being close to N. These tests total more than 10^9 random parameter sets in the range $N \le 10^9 \wedge 10^{-9} \le \omega \le 10^9$.

The rejection method is recommended when σ and n are so high that other sampling methods are inefficient. Since the calculation time for f(x) is quite high, the economy of this method depends mainly on the number of evaluations of f(x) needed for finding the mode and for the subsequent acceptance/rejection loop. The number of function evaluations may be reduced by fast acceptance and fast rejection schemes (Devroye, 1986), using the bounds given by Wallenius (1963):

$$\mathrm{f}_1(x) \le \mathrm{f}(x) \le \mathrm{f}_2(x) \quad \text{for } \omega < 1, \tag{50}$$

$$\mathrm{f}_1(x) \ge \mathrm{f}(x) \ge \mathrm{f}_2(x) \quad \text{for } \omega > 1, \text{ where} \tag{51}$$

$$\mathrm{f}_1(x) = \frac{\Lambda(\mathbf{x})n!}{(m + m_2/\omega)^{\underline{x}}(m_2 + \omega(m-x))^{\underline{x_2}}}, \tag{52}$$

$$\mathrm{f}_2(x) = \frac{\Lambda(\mathbf{x})n!}{(m + (m_2 - x_2)/\omega)^{\underline{x}}(m_2 + \omega m)^{\underline{x_2}}}. \tag{53}$$

The underlined superscript denotes a falling factorial power: $a^{\underline{b}} = a(a-1)...(a-b+1)$. According to (50), we can accept X when $kV < \mathrm{f}_1(\lfloor X \rfloor)$ and reject X when $kV > \mathrm{f}_2(\lfloor X \rfloor)$ for $\omega < 1$ and opposite for

$\omega > 1$. Only when kV lies between these bounds do we need to calculate f(X). The time used for calculating $f_1(X)$ and $f_2(X)$ pays back only when $f_1(x)$ and $f_2(x)$ are close to f(x), which is the case when n is small compared to N, and ω is close to 1.

An alternative improvement, which can eliminate the need for the time-consuming calculation of f(x) in the rejection loop, involves sampling in the τ domain. This method is based on the following theorem:

Let $\mathrm{f}(x) = \int_{\mathrm{T}} \Phi(x,y)dy$ be a the probability function of a distribution on the domain Ξ, where $\Phi(x,y)$ is a non-negative function on the area T. Let the positive constant k and the distribution function h(x) on Ξ be defined so that $\forall x \in \Xi : k\,\mathrm{h}(x) \geq f(x)$. Let $\Upsilon(x,y)$ be a distribution function on the domain T satisfying $\exists x \in \Xi\ \forall y \in \mathrm{T} : \Upsilon(x,y) \geq \dfrac{\Phi(x,y)}{k\,\mathrm{h}(x)}$. For any fixed x for which this condition holds, let Y be a variate in T with the distribution $\Upsilon(x,y)$ and define $q(x,y) = \dfrac{\Phi(x,y)}{k\,\mathrm{h}(x)\Upsilon(x,y)}$. Let the conditional distribution of a variate Z, given Y, be a Bernouilli distribution with parameter $q(x,Y)$. The distribution of Z is then a Bernouilli distribution with parameter $\dfrac{\mathrm{f}(x)}{k\,\mathrm{h}(x)}$, as proven by

$$\Pr(Z=1) = \int_{\mathrm{T}} \Pr(Z=1 \mid Y=y)\Pr(Y=y)dy = \int_{\mathrm{T}} q(x,y)\Upsilon(x,y)dy = \frac{\mathrm{f}(x)}{k\,\mathrm{h}(x)} \qquad \blacksquare$$

When generating a variate X with distribution f(x) using a rejection method based on the dominating function kh(x), we can avoid the calculation of f(x) in the rejection loop for all x for which $q(x,y) \leq 1$, by sampling Y in the T domain from the distribution $\Upsilon(x,y)$ and then sampling Z from the conditional distribution given Y. The number of x-values for which this method is applicable can be increased, if desired, by appropriate choice of k.

To apply this method to Wallenius' distribution, we use the expression $\mathrm{f}(\mathbf{x}) = \Lambda(\mathbf{x})\int_0^1 \Phi_1(\tau)d\tau$, where the integrand $\Phi_1(\tau) = rd\,\tau^{rd-1}\prod_{i=1}^{c}(1-\tau^{r\omega_i})^{x_i}$ has its maximum in τ = ½ when r is the solution to $d - \dfrac{1}{r} - \sum_{i=1}^{c}\dfrac{x_i\omega_i}{2^{r\omega_i}-1} = 0 \ \wedge\ r > \dfrac{1}{d}$, according to Fog (2007). To remove asymmetries, we replace $\Lambda(\mathbf{x})\Phi_1(\tau)$ by $\Phi_2(\mathbf{x},\tau) = \Lambda(\mathbf{x})\{\Phi_1(\tau) + \Phi_1(1-\tau)\}/2$. This function can be approximated by the Gauss curve $\Upsilon_1(\mathbf{x},\tau) = \Phi_2(\mathbf{x},\tfrac{1}{2})\exp\left\{-\dfrac{(\tau-\frac{1}{2})^2}{2w^2}\right\}$, where $w = \dfrac{\ell}{\sqrt{-\varphi''(\frac{1}{2})}}$, $\varphi(\tau) = \log\Phi_1(\tau)$.

ℓ is a correction factor to make sure $q(x,\tau) \leq 1$ for all τ. Υ_1 needs to be normalized by its integral

$$\Upsilon(\mathbf{x},\tau) = \frac{\Upsilon_1(\mathbf{x},\tau)}{G}, \quad G = \int_0^1 \Upsilon_1(\mathbf{x},\tau)dt = \Lambda(\mathbf{x})\Phi_1(\tfrac{1}{2})w\sqrt{2\pi}\,\mathrm{erf}\left(\frac{1}{w\sqrt{8}}\right). \qquad (54)$$

We can sample from this distribution by sampling from a normal distribution and rejecting values outside the interval [0,1]. It is possible to find the minimum value of ℓ in each case, but this would be a waste of computational resources since a higher value of ℓ may be used without making $q(x,\tau) > 1$. The following formula has been found experimentally to give a good fit:

$$\ell = 1 + k_1 (\log E)^{3/2}, \quad E = \frac{1}{d}\sum_{i=1}^{c} \omega_i m_i, \quad k = k_2 \, \mathrm{f}(M), \quad k_1 = 0.0272, \quad k_2 = 1.01. \tag{55}$$

With these choices of parameters, the condition $q(x,\tau) \le 1$ is satisfied for all *x*-values for which $\rho = \frac{G}{k\,\mathrm{h}(x)} \le 1$. This has been verified experimentally for the univariate distribution by testing with 10^9 random parameter sets in the range $N \le 10^9 \wedge 10^{-9} \le \omega \le 10^9$ with the limitation $\left|\tau - \frac{1}{2}\right| \le 30 \cdot w$. Under the assumption that the empirical tests have been sufficient, the method is therefore exact when $w \ge 1/60$. For $w < 1/60$, it can be calculated that the relative error, if any, is certain to be smaller than 10^{-195}. Applied to the univariate Wallenius' distribution with h(*x*) as given above, this algorithm was found to give a substantial reduction in the sampling time. The set-up time is still dominated by the time required to search for the mode.

6. Sampling from the multivariate Wallenius' noncentral hypergeometric distribution

Of the above methods for sampling from Wallenius' distribution, only the simulation of the urn experiment has been applied to the multivariate distribution $\mathrm{f}(\mathbf{x}) = \mathrm{mwnchyp}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega})$. A rejection method for sampling from the multivariate distribution would be useful, but it is difficult to find a good dominating function h(**x**), and universal methods can be quite inefficient (Devroye, 1997; Johnson, 1987).

6.1 Conditional method

The conditional method, as described above for the mfnchyp distribution, can be applied analogously to the multivariate Wallenius' distribution. The equations (32), (34) and (35) are replaced by

$$\Pr(y_{\mathrm{H}} = y) = \mathrm{f}_{\mathrm{H}}(y) \approx \mathrm{wnchyp}(y; n, m_{\mathrm{H}}, N, \omega_{\mathrm{H}} / \omega_{\mathrm{L}}), \tag{56}$$

$$\Pr(\mathbf{x}_{\mathrm{H}} = \mathbf{x} \mid y_{\mathrm{H}} = y) \approx \mathrm{mwnchyp}(\mathbf{x}; y, \mathbf{m}_{\mathrm{H}}, \boldsymbol{\omega}_{\mathrm{H}}), \tag{57}$$

$$\Pr(\mathbf{x}_{\mathrm{L}} = \mathbf{x} \mid y_{\mathrm{H}} = y) \approx \mathrm{mwnchyp}(\mathbf{x}; n - y, \mathbf{m}_{\mathrm{L}}, \boldsymbol{\omega}_{\mathrm{L}}). \tag{58}$$

Intuitively, we would expect (57) and (58) to be exact, just like (34) and (35) are, but unfortunately this is not the case. Nevertheless, (57) and (58) are quite good approximations in most cases, so that the precision of this method is determined predominantly by (56). Experiments show that a reasonable precision is obtained when differences in weight within groups is not too large.

6.2 Gibbs sampling

While the Gibbs sampler for the multivariate Fisher's noncentral hypergeometric distribution is exact in the limit, the analogous sampler for the multivariate Wallenius' distribution is not, because the conditional distribution is only approximately a univariate Wallenius' distribution, and the exact conditional distribution is not simple to calculate. The best convergence is obtained by sorting the colors by variance, but the best accuracy is obtained by sorting the colors by weight and skipping the wrap-around steps where $i = c$ and $j = 1$. This method generally gives better accuracy than the conditional method alone.

6.3 Metropolis-Hastings sampling

A Metropolis-Hastings sampler is a Markov chain where the transition from state **x** to state **y** is defined by the proposal density h(**y**;**x**) and the probability of acceptance p(**y**;**x**). Each step consists of drawing a

new proposed state **y** from h and accepting it with probability p. If **y** is not accepted then **x** is retained (Hastings, 1970). The limiting distribution f(**x**) is approached if the following reversibility condition is satisfied

$$\frac{\mathrm{p}(\mathbf{y};\mathbf{x})}{\mathrm{p}(\mathbf{x};\mathbf{y})} = \frac{\mathrm{f}(\mathbf{y})\,\mathrm{h}(\mathbf{x};\mathbf{y})}{\mathrm{f}(\mathbf{x})\,\mathrm{h}(\mathbf{y};\mathbf{x})}. \tag{59}$$

In order to maximize the acceptance rate, we choose for p the Metropolis dynamics (Peskun, 1973):

$$\mathrm{p}(\mathbf{y};\mathbf{x}) = \min\left\{1, \frac{\mathrm{f}(\mathbf{y})\,\mathrm{h}(\mathbf{x};\mathbf{y})}{\mathrm{f}(\mathbf{x})\,\mathrm{h}(\mathbf{y};\mathbf{x})}\right\}, \tag{60}$$

which satisfies (59). When f is a multivariate distribution of *c* dimensions, we can define a scan as a series of *c* consecutive steps in one dimension each, as in the Gibbs sampler. If the proposal distribution for a step in dimension *i* is the conditional distribution (36) then the acceptance rate is 1 and the Metropolis-Hastings sampler is reduced to a Gibbs sampler. In the case of the multivariate Wallenius' noncentral hypergeometric distribution, we can choose the univariate Wallenius' distribution as the proposal density. For a step in color *i* with color *j* as the dependent

$$\mathrm{h}(y_i; x_i) = \mathrm{wnchypg}(y_i; x_i + x_j, m_i, m_i + m_j, \omega_i / \omega_j), \quad y_j = x_i + x_j - y_i. \tag{61}$$

This is such a good approximation to the conditional distribution that the rate of acceptance will be close to 1 and the convergence will be almost as good as for a Gibbs sampler. The conditions for convergence are aperiodicity and communication between all points (Smith and Roberts, 1993). These conditions are met if all m_i and ω_i are positive. As before, it is preferred to use the approximate variate obtained by the conditional method as a starting point and to sort the colors by variance. We do not have an accurate way of calculating the variance of the multivariate Wallenius' distribution, but we may use the coarse approximation obtained by approximating the mwnchypg distribution with a mfnchypg distribution with the same mean and using equation (38). Potential minor errors in the sort order caused by this inaccuracy have little influence on the performance. The method is still exact in the limit. This method is fairly time consuming, but no other known method gives the same precision faster when *n* is high. We may replace the proposal distribution given by (61) with another distribution such as fnchypg. Sampling from fnchypg is considerably faster than sampling from wnchypg, but the imprecision introduced hereby reduces the acceptance rate so that a longer burn-in period is needed.

6.4 Dividing into subsamples

The accuracy of any of the preceding methods for mwnchypg can be improved by dividing *n* into *S* smaller samples of positive size n_j, so that $\sum_{j=1}^{S} n_j = n$. Generate a series of variates $\mathbf{Y}_j$ (j = 1 .. S) that each approximate a multivariate Wallenius noncentral hypergeometric distribution with the parameters $\mathbf{Y}_j \sim \mathrm{mwnchyp}(n, \mathbf{m} - \sum_{k=1}^{j-1} \mathbf{Y}_k, \boldsymbol{\omega})$. The combined sample $\mathbf{X} = \sum_{j=1}^{S} \mathbf{Y}_j$ will then approach the distribution $\mathrm{mwnchyp}(n, \mathbf{m}, \boldsymbol{\omega})$ when $S \to n$ if the distribution of subsamples $\mathbf{Y}_j$ approach exactness when $n_j \to 1$, because the taking of *n* subsamples of size 1 is identical to the urn experiment. This method is not very efficient, however, because *S* has to be quite high in order to obtain a good accuracy.

7. Suggestions for future research

There is a need for more efficient and exact sampling methods for the multivariate distributions, possibly rejection methods or conditional methods. Theoretically justified hat functions for the rejection methods

may be more satisfying than the experimentally obtained formulas.

REFERENCES

Ahrens, J. H. and Dieter, U. (1989). A convenient sampling method with bounded computation times for Poisson distributions. *American Journal of Mathematical and Management Sciences* **9**, 1-13.

Besag, J., Green, P., Higdon, D., and Mengerson, K. (1995). Bayesian Computation and Stochastic Systems. *Statistical Science* **10**, 3-66.

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *American Statistician* **46**, 167-174.

Cheng, R. C. H. (1998). Random Variate Generation. In *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, Banks, J. (ed), 139-172. New York: John Wiley & Sons.

Chesson, J. (1976). A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. *Journal of Applied Probability* **13**, 795-797.

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.

Devroye, L. (1997). Random variate Generation for Multivariate Unimodal Densities. *ACM transactions on Modeling and Computer Simulation* **7**, 447-477.

Fog, A. (2007). Calculation methods for Wallenius' noncentral hypergeometric distribution. Working paper. www.agner.org/random/theory/nchyp1.pdf.

Harkness, W. L. (1965). Properties of the Extended Hypergeometric Distribution. *Annals of Mathematical Statistics* **36**, 938-945.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97-109.

Johnson, M. E. (1987). *Multivariate Statistical Simulation.* New York: Wiley & Sons.

Levin, B. (1984). Simple improvements on Cornfield's approximation to the mean of a noncentral hypergeometric random variable. *Biometrika* **71**, 630-632.

Liao, J. (1992). An Algorithm for the Mean and Variance of the Noncentral Hypergeometric Distribution. *Biometrics* **48**, 889-892.

Liao, J. G. and Rosen O. (2001). Fast and Stable Algorithms for Computing and Sampling from the Noncentral Hypergeometric Distribution. The American Statistician **55**, 366-369.

Manly, B. F. J. (1974). A Model for Certain Types of Selection Experiments. *Biometrics* **30**, 281-294.

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

McDonald, J. W., Smith, P. W. F. and Forster, J. J. (1999). Exact Tests of Goodness of Fit of Log-Linear Models for Rates. *Biometrics* **55**, 620-624.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607-612.

Roberts, G. O. and Polson, N. G. (1994). On the Geometric Convergence of the Gibbs Sampler. *Journal of the Royal Statistical Society, ser. B.* **56**, 377-384.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, ser. B.* **55**, 3-23.

Stadlober, E. (1989). Sampling from poisson, binomial and hypergeometric distributions: Ratio of uniforms as a simple and fast alternative. *Berichte der mathematisch-statistischen Sektion in der Forschungsgesellschaft Joanneum*, no. 303, Graz.

Stadlober, E. (1990). The ratio of uniforms approach for generating discrete random variates. *Journal of Computational and Applied Mathematics* **31**, 181-189.

Wallenius, K. T. (1963). *Biased Sampling: The Non-central Hypergeometric Probability Distribution*. Ph.D. thesis, Stanford University (Also published with the same title as Technical report no. 70), Department of Statistics, Stanford University.