Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution

Agner Fog, 2008-02-06.

A revised version of this article is published in *Communications in Statistics, Simulation and Computation*, vol. 37, no. 2, pp. 258-273, 2008.

SUMMARY. Two different probability distributions are both known in the literature as "the" noncentral hypergeometric distribution. The nomenclature problems are discussed. Wallenius' noncentral hypergeometric distribution can be described as an urn model with bias. Fisher's noncentral hypergeometric distribution is the conditional distribution of independent binomial variates given their sum. Several different methods for calculating probabilities from Wallenius' noncentral hypergeometric distribution are derived. Range of applicability, numerical problems and efficiency are discussed for each method. Approximations to the mean and variance are also discussed. This distribution has important applications in models of biased sampling and in models of evolutionary systems.

KEY WORDS: Noncentral hypergeometric distribution; Wallenius; Fisher; Multivariate distribution; Probability function.

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. The nomenclature problems are discussed below. Fisher's noncentral hypergeometric distribution is the conditional distribution of independent binomial variates given their sum (McCullagh and Nelder, 1983). Wallenius' noncentral hypergeometric distribution is a distribution of biased sampling. It can be described as an urn model without replacement with bias. Wallenius' distribution has many potential applications including models of selective survival and selective predation in ecology and evolutionary biology (Manly, 1985), as well as general models of biased sampling (Graves and Hamada, 2006; Wallenius, 1963). The application of this distribution has been hampered by the fact that the only published calculation method (Lyons, 1980) is numerically unstable, inefficient, and applicable only to a narrow range of parameters, as explained below. The purpose of the present study is to seek reliable calculation methods that are applicable to a wide range of parameters, including the multivariate case. Methods for sampling from this distribution are described in an accompanying paper (Fog, 2007).

2. Definition and properties

Assume that an urn contains $N = \sum_{i=1}^{c} m_i$ balls of c different colors, where m_i is the number of balls of color $i \in C = \{1,...,c\}$. n balls are sampled, one by one, from the urn without replacement in such a way that the probability that a particular ball is sampled at a given draw is proportional to a property ω_i which we will call weight or odds. The weight of a ball depends only on its color i. Let $\mathbf{X}_v = (X_{1v}, X_{2v},..., X_{cv})$ denote the total number of balls of each color sampled in the first v draws. The probability that the next draw gives a ball of color i is

$$p_{i(\nu+1)}(\mathbf{X}_\nu) = \frac{(m_i - X_{i\nu})\omega_i}{\sum_{j=1}^{c}(m_j - X_{j\nu})\omega_j}. \tag{1}$$

The probability function for this distribution has been derived by Wallenius (1963) for the univariate case ($c = 2$) and by Chesson (1976) for the multivariate case:

$$\text{mwnchypg}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) = \Lambda(\mathbf{x})\mathrm{I}(\mathbf{x}), \text{ where} \tag{2}$$

$$\Lambda(\mathbf{x}) = \prod_{i=1}^{c}\binom{m_i}{x_i}, \tag{3}$$

$$\mathrm{I}(\mathbf{x}) = \int_0^1 \prod_{i=1}^{c}(1 - t^{\omega_i/d})^{x_i}\, dt, \tag{4}$$

$$d = \boldsymbol{\omega} \cdot (\mathbf{m} - \mathbf{x}) = \sum_{i=1}^{c}\omega_i(m_i - x_i), \tag{5}$$

$$\mathbf{x} = (x_1, x_2, ..., x_c), \quad \mathbf{m} = (m_1, m_2, ..., m_c), \ \boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_c), \tag{6}$$

which is valid for $d > 0$. The unexpected integral in (4) arises as the solution to a discrete difference equation (Wallenius, 1963; Chesson, 1976).

The odds can be arbitrarily scaled:

$$\forall r > 0 : \text{mwnchypg}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) = \text{mwnchypg}(\mathbf{x}; n, \mathbf{m}, r\boldsymbol{\omega}). \tag{7}$$

The univariate distribution ($c = 2$) can be defined as the probability function

$$\text{wnchypg}(x; n, m, N, \omega) = \text{mwnchypg}\{(x_1, x_2); n, (m_1, m_2), (\omega_1, \omega_2)\}, \tag{8}$$

where $x_1 = x$, $x_2 = n\text{-}x$, $m_1 = m$, $m_2 = N\text{-}m$, $\omega_1 = \omega$, $\omega_2 = 1$. The following properties of the univariate distribution are easily derived:

$$\text{wnchypg}(x; n, m, N, \omega) = \text{wnchypg}(n - x; n, N - m, N, 1/\omega), \ \omega \neq 0. \tag{9}$$

$$\begin{aligned}\text{wnchypg}(x; n, m, N, \omega) = {} & \text{wnchypg}(x-1; n-1, m, N, \omega)\frac{(m - x + 1)\omega}{(m - x + 1)\omega + N - n - m + x} + \\ & \text{wnchypg}(x; n-1, m, N, \omega)\frac{N - n - m + x + 1}{(m - x)\omega + N - n - m + x + 1}.\end{aligned} \tag{10}$$

$$\begin{aligned}\text{wnchypg}(x; n, m, N, \omega) = {} & \text{wnchypg}(x-1; n-1, m-1, N-1, \omega)\frac{m\omega}{m\omega + m_2} + \\ & \text{wnchypg}(x; n-1, m, N-1, \omega)\frac{m_2}{m\omega + m_2}.\end{aligned} \tag{11}$$

The distribution of the balls that are left in the urn is not a Wallenius' noncentral hypergeometric distribution. This is a lack of symmetry that distinguishes Wallenius' from Fisher's noncentral

hypergeometric distribution. We will therefore define the complementary Wallenius' noncentral hypergeometric distribution as the distribution of the balls that remain in the urn:

$$\text{mcwnchypg}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) = \text{mwnchypg}\left\{\mathbf{m}-\mathbf{x}; N-n, \mathbf{m}, \left(\frac{1}{\omega_1}, ..., \frac{1}{\omega_c}\right)\right\}, \ \omega_i > 0. \qquad (12)$$

This function is used for modeling the distribution of survivors after a Darwinian selection process (Manly, 1974).

3. Nomenclature problem

The two noncentral hypergeometric distributions are often confused in the literature or falsely assumed to be identical (e.g. Lyons, 1980; SAS Institute, 2002). Therefore I decided in 2003 to find a solution to the name conflict. The history of the nomenclature, as far as I was able to trace it, was as follows:

Wallenius' distribution was given the name noncentral hypergeometric distribution in 1963 by K.T. Wallenius. Manly (1972, 1974, 1985) has used this distribution without knowing about the preceding literature and without knowing that the distribution had a name.

Fisher's distribution was first described by R. A. Fisher (1935) who did not give it a name. It was given the name extended hypergeometric distribution in 1965 by W. L. Harkness. This name is rarely used, while the name noncentral hypergeometric for Fisher's distribution is the prevalent name in the literature today. Wallenius and Harkness have actually met each other around 1965 and discussed the two distributions. According to Wallenius' recollection, Harkness commented, "I wish I had come up with the name noncentral hypergeometric".

The use of the name noncentral hypergeometric for Fisher's distribution can be traced back to the following publications: Breslow and Day (1980), McCullagh and Nelder (1983), Levin (1984). I have contacted these authors in order to trace the origin of this name and solicit their opinion on the naming problem. Norman Breslow, Nick Day and Peter McCullagh cannot remember where they have the name from. In their 1980 book, Breslow and Day wrote that this distribution "is known in the probability literature as the noncentral hypergeometric distribution". In a reply to me, Breslow apologizes for the name confusion and explains, "I learned about this distribution from the papers by Hannan and Harkness (1963) and Gart but see that they did not use this terminology". Bruce Levin has the clearest recollection. He refers to several articles by J. J. Gart (1962 and later) who calls ω a parameter of non-centrality. Levin considers the name noncentral hypergeometric distribution an obvious extension to Gart's terminology. So obvious, in fact, that he "never thought twice about it".

According to the replies I have received from the abovementioned authors, there is no evidence that they have a common origin for the use of the name noncentral for Fisher's distribution, though there may be one. It is equally possible that they have all constructed the name from an obvious analogy with other noncentral distributions under the influence of Gart's "parameter of non-centrality". A third possibility is that one or more of these authors have heard the name noncentral hypergeometric in connection with Wallenius' distribution and assumed that Fisher's distribution was meant or that the two distributions were identical.

Common reference handbooks use the name extended hypergeometric for Fisher's distribution (Johnson et al. 1969, 1992; Marriott 1990) and so does common commercial software (SAS Institute 2002). Johnson et al. (1969, 1992, 1997) use the name noncentral hypergeometric for

Wallenius' distribution.

Discussions with all the relevant scientists I could trace, has led me to the conclusion that the best solution to the name conflict is to apply the prefixes Wallenius' and Fisher's to the name in order to distinguish the two distributions. Some of my correspondents were strongly opposed to using the name extended hypergeometric for Fisher's distribution. While the use of prefixes 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.

(I have corresponded with the following scientists about this issue: Norman Breslow, Jean Chesson, Nick Day, James Gentle, Carlos Hernandez-Suarez, Norman L. Johnson, Samuel Kotz, Bruce Levin, Jiangang Liao, Bryan Manly, Peter McCullagh, Jordi Ocaña, and Ted Wallenius).

4. Mean and variance

The mean $\boldsymbol{\mu}_\nu = (\mu_{1\nu},...,\mu_{c\nu})$ of $\mathbf{X}_\nu$ in Wallenius' noncentral hypergeometric distribution can be approximated by

$$\mu_{i\nu} \approx \mu_{i(\nu-1)} + p_{i\nu}(\boldsymbol{\mu}_{\nu-1}) \ , \quad \mu_{i0} = 0 \, . \tag{13}$$

This set of difference equations can be approximated by a set of differential equations with the solution given by (Manly, 1974):

$$\left(1-\frac{\mu_1^*}{m_1}\right)^{1/\omega_1} = \left(1-\frac{\mu_2^*}{m_2}\right)^{1/\omega_2} = ... = \left(1-\frac{\mu_c^*}{m_c}\right)^{1/\omega_c} \quad \wedge \tag{14}$$

$$\sum_{i=1}^{c} \mu_i^* = n \ \wedge \ \forall i \in C : 0 \le \mu_i^* \le m_i \, .$$

The solution $\boldsymbol{\mu}^* = (\mu_1^*,...,\mu_c^*)$ is an approximation to the mean $\boldsymbol{\mu}$ of $\mathbf{X}$, which is valid under the conditions that $\forall i \in C : m_i > 0 \wedge \omega_i > 0$. Interestingly, the mean given by (14) is a good approximation, and in most cases better than the value obtained by iteration of (13). (14) is exact when all ω_i are equal, while (13) is not. Manly (1974) derived (14) for the purpose of estimating the ω_i's from experimental samples. To solve (14) for the means, define

$$e^{\theta} = \left(1-\frac{\mu_i^*}{m_i}\right)^{1/\omega_i} \ , \quad z(\theta) = \sum_{i=1}^{c} m_i\left(1-e^{\omega_i\theta}\right) \tag{15}$$

and solve $z(\theta) = n$ by Newton-Raphson iteration. In the univariate case with $\omega > 1$, it may be more efficient to solve

$$\left(1-\frac{n-\mu^*}{N-m}\right)^{\omega} - 1 + \frac{\mu^*}{m} = 0 \, . \tag{16}$$

Manly, Miller and Cook (1972) give an exact expression for the variance σ^2 of the univariate distribution as a function of $\mu_{1\nu}$, $\nu = 1,...,n$. Unfortunately, this expression is so sensitive to inaccuracies in the means $\mu_{1\nu}$, that it is useless unless all the means are known with very high precision.

An approximation to the variance can be obtained by approximating Wallenius' noncentral hypergeometric distribution with a Fisher's noncentral hypergeometric distribution with the same mean and using an approximate formula given by Levin (1984) for the variance of the latter distribution:

$$\sigma^2 \approx \sigma_F^2 = \frac{Nab}{(N-1)(mb+(N-m)a)},$$
$$a = \mu^*(m-\mu^*),\ b = (n-\mu^*)(\mu^*+N-n-m). \qquad (17)$$

This approximation is good when ω is near 1 and n is far from N.

A simple relationship between σ and the maximum of the probability function f is obtained from the normal distribution approximation:

$$\sigma \approx \sigma_N = \frac{1}{\mathrm{f}(M)\sqrt{2\pi}}, \qquad (18)$$

where M is the mode. This approximation is better than σ_F in some cases where σ is high. However, (18) has an obvious limitation since $\sigma_N \geq 1/\sqrt{2\pi}$.

5. Methods for calculation of the probability function

Only one calculation method has hitherto been mentioned in the literature (Lyons, 1980). Unfortunately, this method is inefficient and entails serious numerical problems, as explained below. Several other methods will be developed here, and the applicability of each method will be discussed.

5.1 Recursive calculation

The most obvious calculation method for the univariate distribution is the recursive application of (10). Figure 1 illustrates this method. The field with coordinates (ν,ξ) represents the probability that there are exactly ξ balls of color 1 among the first ν balls drawn from the urn. The value of each field is calculated from the field to the left of it and the field below the latter. The chain of arrows constitutes an arbitrary trajectory. This method has excellent numerical stability for all parameter values. The accumulation of rounding errors is not severe. Numerical underflow can occur, but can safely be ignored. This method is inefficient when n and x are high, because the number of probabilities to calculate is $n(x+1)-x^2$. The economy of this method can be improved by ignoring negligible probabilities far from the mean. The recursive method can be used in all situations where the economy of computer resources allows it. This method becomes complicated and inefficient in the multivariate case.

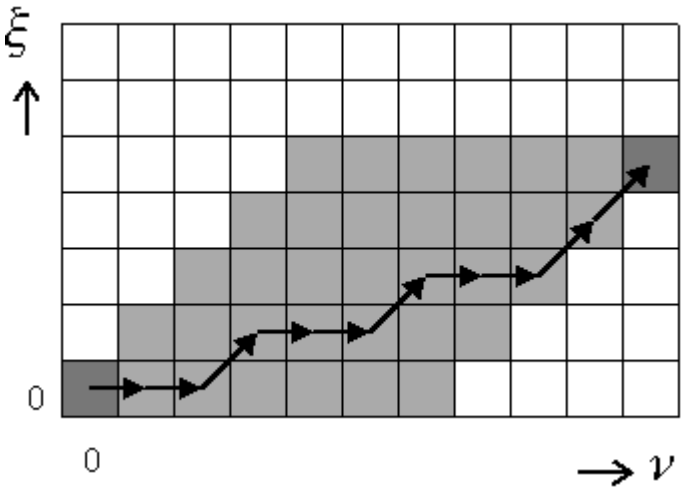

fig. 1. Recursive calculation of Wallenius probabilities.

5.2 Binomial expansion methods

Let r be a positive scale factor, and substitute $t = \tau^{rd}$ into (4):

$$\mathrm{I}(\mathbf{x}) = rd\int_0^1 \tau^{rd-1}\prod_{i=1}^{c}(1-\tau^{r\omega_i})^{x_i}\,d\tau\,. \tag{19}$$

Consider the univariate case ($c = 2$), let $\omega_1 = \omega$, $\omega_2 = 1$, $r = 1$, and apply (2) and (3):

$$\mathrm{wnchypg}(x;n,m,N,\omega) = \binom{m}{x}\binom{m_2}{x_2}d\int_0^1 (1-\tau^{\omega})^{x}(1-\tau)^{x_2}\tau^{d-1}d\tau\,. \tag{20}$$

Applying the binomial theorem to $(1-\tau^{\omega})^{x}$ and swapping the order of integration and summation gives:

$$\mathrm{wnchypg}(x;n,m,N,\omega) = \binom{m}{x}\binom{m_2}{x_2}d\sum_{j=0}^{x}(-1)^j\binom{x}{j}\int_0^1 (1-\tau)^{x_2}\tau^{d+\omega j-1}d\tau\,. \tag{21}$$

This integral is known as the Beta function

$$\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{22}$$

hence

$$\mathrm{wnchypg}(x;n,m,N,\omega) = \binom{m}{x}\binom{m_2}{x_2}d\sum_{j=0}^{x}(-1)^j\binom{x}{j}\mathrm{B}(x_2+1,d+\omega j)\,. \tag{23}$$

Using $\Gamma(a+1) = a!$, this can be reduced to

$$\mathrm{wnchypg}(x;n,m,N,\omega) = m^{\underline{x}}m_2^{\underline{x_2}}d\sum_{j=0}^{x}q_{jx}\,,\quad q_{jx} = \frac{(-1)^j}{j!(x-j)!(\omega(m+j-x)+m_2)^{\underline{x_2+1}}}\,, \tag{24}$$

where the notation $a^{\underline{b}}$ means the falling factorial power, defined by

$$a^{\underline{b}} = \prod_{k=a-b+1}^{a} k = \frac{\Gamma(a+1)}{\Gamma(a-b+1)}. \tag{25}$$

The following recursion formula holds for $j>0 \wedge x>0$:

$$q_{jx} = -q_{(j-1)(x-1)} \frac{\omega(m - x + j) + m_2 - x_2 - 1}{j}. \tag{26}$$

As a corollary of (24), the case $x = 0$ gives

$$\text{wnchypg}(0; n, m, N, \omega) = \frac{m_2^{\underline{n}}}{(m_2 + \omega m)^{\underline{n}}}. \tag{27}$$

This method can be expanded to the multivariate case by binomial expansion of the powers in (19) for all but the largest of x_i, giving c-1 nested sums. In the case where all but one of the x_i values are zero, we get

$$\text{mwnchypg}\{(0,0,...,x_j,0,0,..); n, \mathbf{m}, \boldsymbol{\omega}\} = \frac{m_j^{\underline{n}}}{\left(\frac{1}{\omega_j}\sum_{i=1}^{c} m_i\omega_i\right)^{\underline{n}}}. \tag{28}$$

The summation in (24) can cause serious numerical problems, even for moderate values of the parameters. For example, a calculation of wnchypg(46;80,50,100,5) using (24) with double precision gives the value -34.49. The correct value is 0.002530. This error is due to loss of precision, evidenced by the fact that the numerically largest of the q_{jx} terms is $1.3 \cdot 10^{19}$ times as large as than the final sum. This method is therefore not reliable unless the number of summation terms is quite small.

The method given by Lyons (1980) is obtained by binomial expansion of both powers in (4) for the univariate case. This gives

$$\text{wnchyp}(x; n, m, N, \omega) = \binom{m}{x}\binom{m_2}{x_2}\sum_{j=0}^{x}\left\{\binom{x}{j}\sum_{k=0}^{x_2}\binom{x_2}{k}\frac{(-1)^{j+k} d}{k + \omega j + d}\right\}. \tag{29}$$

The numerical problems in (29) are generally worse than in (24). A calculation of the same numerical example as above with this method gives the somewhat dubious result $-6.04 \cdot 10^{22}$. The error is due to loss of precision in both sums. For the outer sum, the numerically largest term is $7.2 \cdot 10^{26}$ times the final sum. Furthermore, (29) is less economical than (24) because of the nested sums. Lyons' method is therefore not recommendable.

5.3 Taylor expansion methods

Consider the definite integral of an arbitrary function:

$$I(\delta) = \int_{\tau_0-\delta}^{\tau_0+\delta} \Phi(\tau) d\tau. \tag{30}$$

Define an auxiliary function $\Upsilon(\tau)$ and a correction function $\Psi(\tau) = \Phi(\tau)/\Upsilon(\tau)$. Assume that these three functions are all analytic in the complex disk $\{\tau \in \mathbb{C} \,|\, |\tau - \tau_0| < R\}$. Expanding $\Psi(\tau)$ in a Taylor series and swapping the order of integration and summation gives

$$I(\delta) = \sum_{j=0}^{\infty} \int_{\tau_0-\delta}^{\tau_0+\delta} \Upsilon(\tau) \frac{\Psi^{(j)}(\tau_0)}{j!} (\tau - \tau_0)^j \, d\tau \,. \tag{31}$$

This expansion is convergent for $\delta < R$ because integration does not change the radius of convergence. If $\Upsilon(\tau)$ is symmetric around τ_0 then the odd terms in the sum vanish:

$$\begin{aligned} I(\delta) &= \sum_{j=0}^{\infty} \int_{\tau_0-\delta}^{\tau_0+\delta} \Upsilon(\tau) \frac{\Psi^{(2j)}(\tau_0)}{(2j)!} (\tau - \tau_0)^{2j} \, d\tau \\ &= 2 \sum_{j=0}^{\infty} \int_{\tau_0}^{\tau_0+\delta} \Upsilon(\tau) \frac{\Psi^{(2j)}(\tau_0)}{(2j)!} (\tau - \tau_0)^{2j} \, d\tau. \end{aligned} \tag{32}$$

The auxiliary function $\Upsilon(\tau)$ should be chosen so that this integral can be calculated analytically. The convergence of this expansion is most likely to be good if $\Upsilon(\tau)$ is chosen so that $\Psi(\tau)$ has most of its weight near τ_0.

A Taylor method is not suited for the integral (4) because the integrand has most of its weight near 0 where, in most cases, it is not differentiable. The transformed integral (19) is preferred. Let

$$\Phi(\tau) = rd\tau^{rd-1} \prod_{i=1}^{c} (1 - \tau^{r\omega_i})^{x_i} \,. \tag{33}$$

Logarithmating and differentiating gives

$$\varphi(\tau) = \log \Phi(\tau) = \log(rd) + (rd-1)\log(\tau) + \sum_{i=1}^{c} x_i \log(1 - \tau^{r\omega_i}) \,, \tag{34}$$

$$\varphi'(\tau) = \frac{rd-1}{\tau} - \sum_{i=1}^{c} \frac{x_i r \omega_i \tau^{r\omega_i - 1}}{1 - \tau^{r\omega_i}} \,. \tag{35}$$

Analysis of $\tau\varphi'(\tau)$ shows that $\varphi(\tau)$ and $\Phi(\tau)$ have a single maximum in the interval $0 < \tau_0 < 1$ when $r > 1/d$. The preferred value of the mode τ_0 is ½. In order to obtain this value, we define

$$z(r) = \frac{\varphi'(\frac{1}{2})}{2r} = d - \frac{1}{r} - \sum_{i=1}^{c} \frac{x_i \omega_i}{2^{r\omega_i} - 1} \,, \tag{36}$$

and observe that the equation $z(r) = 0$ has a unique solution in the interval $1/d < r < \infty$, which can be found by the Newton-Raphson iteration:

$$r_{j+1} = \left\{ \begin{array}{ll} r_j - \dfrac{z(r_j)}{z'(r_j)} & \text{if this is} > \dfrac{1}{d} \\ \dfrac{r_j + 7}{8d} & \text{otherwise} \end{array} \right\}, \quad z'(r) = \frac{1}{r^2} + \sum_{i=1}^{c} \frac{x_i \omega_i^{\,2} 2^{r\omega_i} \log(2)}{(2^{r\omega_i} - 1)^2} \,. \tag{37}$$

Henceforth, we will assume that r is the solution to $z(r) = 0$, so that the mode $\tau_0 = ½$. With these values of r and τ_0, we can apply (32) to the calculation of the integral (19).

Three different choices for the auxiliary function $\Upsilon(\tau)$ will be explored:

$$\Upsilon_1(\tau) = 1, \tag{38}$$

$$\Upsilon_2(\tau) = A_0 e^{a_2\left(\tau-\frac{1}{2}\right)^2}, \quad \mathrm{A}_0 = \Phi(\tfrac{1}{2}), \quad a_2 = \tfrac{1}{2}\varphi''(\tfrac{1}{2}), \tag{39}$$

$$\Upsilon_3(\tau) = A_0\{4\tau(1-\tau)\}^b, \quad b = -\tfrac{1}{8}\varphi''(\tfrac{1}{2}). \tag{40}$$

The parameters of Υ_2 and Υ_3 are both chosen so that the first two derivatives are equal to the derivatives of Φ. The first choice, $\Upsilon(\tau) = \Upsilon_1(\tau)$, gives

$$I(\delta) = \sum_{j=0}^{\infty} \frac{2\Phi^{(2j)}(\frac{1}{2})}{(2j+1)!}\delta^{2j+1}. \tag{41}$$

The derivatives $\Phi^{(k)}(½)$ are obtained by logarithmic differentiation. Applying Leibniz's rule for differentiation of a product to $\Phi'(\tau) = \Phi(\tau)\varphi'(\tau)$ gives

$$\Phi^{(k)}(\tau) = \sum_{j=1}^{k}\binom{k-1}{j-1}\Phi^{(k-j)}(\tau)\varphi^{(j)}(\tau). \tag{42}$$

The k'th derivative of φ for $k > 0$ can be expressed by the formula

$$\varphi^{(k)}(\tau) = \frac{(-1)^{k-1}(k-1)!(rd-1)}{\tau^k} - \sum_{i=1}^{c}\sum_{j=1}^{k}\frac{x_i\zeta_{ijk}\tau^{jr\omega_i-k}}{(1-\tau^{r\omega_i})^j}, \tag{43}$$

$$\zeta_{ijk} = \begin{Bmatrix} \zeta_{ij(k-1)}(jr\omega_i - k + 1) + \zeta_{i(j-1)(k-1)}(j-1)r\omega_i & \text{for} & k > 1 \wedge 0 < j \le k \\ r\omega_i & \text{for} & k = j = 1 \\ 0 & \text{otherwise} & \end{Bmatrix},$$

or alternatively for k > 1:

$$\varphi^{(k)}(\tau) = \frac{1-k}{\tau}\varphi^{(k-1)}(\tau) - \sum_{i=1}^{c}\sum_{j=1}^{k}\frac{x_i\eta_{ijk}\tau^{jr\omega_i-k}}{(1-\tau^{r\omega_i})^j}, \tag{44}$$

$$\eta_{ijk} = \begin{Bmatrix} \eta_{ij(k-1)}(jr\omega_i - k + 2) + \eta_{i(j-1)(k-1)}(j-1)r\omega_i & \text{for} & k > 1 \wedge 0 < j \le k \\ r\omega_i & \text{for} & k = j = 1 \\ 0 & \text{otherwise} & \end{Bmatrix}.$$

(43) and (44) are both proved by induction.

The nearest singularities of $\Phi(\tau)$ are $\tau = 0$ and 1 (except in the rare case that all powers are integers). Therefore, the radius of convergence is ½, and the expansion (41) is convergent for $\delta <$ ½. A typical shape of the integrand curve $\Phi(\tau)$ is shown in fig. 2a. In order to estimate the relative error when the integration interval is narrowed to $\delta <$ ½, we approximate the integrand

curve $\Phi(\tau)$ with the Gauss curve $\Upsilon_2(\tau)$, given by (39). The relative error due to the narrowed integration interval is

$$\varepsilon = \frac{1}{\mathbf{I(x)}}\int_0^{\frac{1}{2}-\delta}\{\Phi(\tau)+\Phi(1-\tau)\}d\tau \approx \int_{-\frac{1}{2}}^{-\delta} e^{a_2\tau^2}d\tau \Big/ \int_{-\frac{1}{2}}^{0} e^{a_2\tau^2}d\tau < \int_{-\infty}^{-\delta} e^{a_2\tau^2}d\tau \Big/ \int_{-\infty}^{0} e^{a_2\tau^2}d\tau = \operatorname{erfc}(\delta\sqrt{-a_2}) \ . \tag{45}$$

Thus, we can find a suitable value of δ from the desired precision ε:

$$\delta = \min\left\{\tfrac{1}{2}, \frac{\operatorname{erfc}^{-1}(\varepsilon)}{\sqrt{-a_2}}\right\}. \tag{46}$$

The atypical case in fig. 2b, where the Gauss curve is a poor approximation to Φ, will lead to an overestimation of the relative error ε and to the conclusion that the Taylor method is unsuitable.

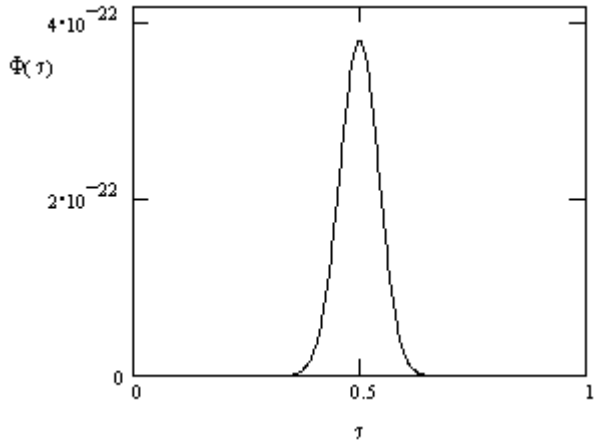

fig. 2a. Typical shape of integrand $\Phi(\tau)$. (x=40, n=80, m=50, N=100, ω=2, c=2)

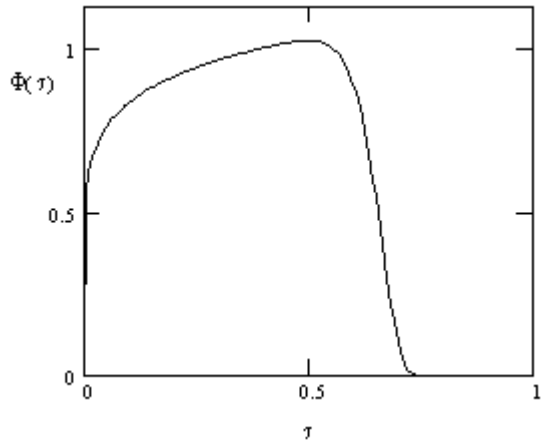

fig. 2b. Integrand with distorted shape. (x=999, n=999, m=999, N=1000, ω=15, c=2)

These results regarding convergence, precision, and the value of δ also pertain to the second and third choice for $\Upsilon(\tau)$ discussed below.

The second choice, $\Upsilon(\tau) = \Upsilon_2(\tau)$, gives

$$I(\delta) = A_0\sum_{j=0}^{\infty}\frac{\Psi^{(2j)}(\frac{1}{2})}{(2j)!}\int_{\frac{1}{2}-\delta}^{\frac{1}{2}+\delta} e^{a_2(\tau-\frac{1}{2})^2}\left(\tau-\tfrac{1}{2}\right)^{2j}d\tau \ . \tag{47}$$

Since $a_2 < 0$ we can define $v = \frac{1}{2}\sqrt{-a_2}$ and substitute $\theta = 2v(\tau-\frac{1}{2})$:

$$I(\delta) = A_0\sum_{j=0}^{\infty}\frac{\Psi^{(2j)}(\frac{1}{2})}{(2v)^{2j+1}(2j)!}\int_{-2v\delta}^{2v\delta}\theta^{2j}e^{-\theta^2}d\theta \ . \tag{48}$$

This integral resembles the repeated integral of the error function (Abramowitz and Stegun, 1965). Integrating by parts j times, noting that $\int te^{-t^2}dt = -\frac{1}{2}e^{-t^2}$, we have

$$\begin{aligned} G(D, j) &= \int_{-D}^{D} t^{2j} e^{-t^2} dt \\ &= \frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})} \int_{-D}^{D} e^{-t^2} dt - \sum_{k=1}^{j} \left[\frac{\Gamma(j+\frac{1}{2})}{2\Gamma(k+\frac{1}{2})} t^{2k-1} e^{-t^2} \right]_{t=-D}^{t=D} \\ &= \Gamma(j+\tfrac{1}{2}) \left\{ \operatorname{erf}(D) - e^{-D^2} \sum_{k=1}^{j} \frac{D^{2k-1}}{\Gamma(k+\frac{1}{2})} \right\}. \end{aligned} \tag{49}$$

The bracketed expression is a residue of the known series

$$\operatorname{erf}(D) = e^{-D^2} \sum_{k=1}^{\infty} \frac{D^{2k-1}}{\Gamma(k+\frac{1}{2})}. \tag{50}$$

Using $\Gamma(k+\frac{1}{2}) = 2^{-k}(2k-1)!!\sqrt{\pi}$, we can calculate

$$G(D, j) = \Gamma(j+\tfrac{1}{2}) e^{-D^2} \sum_{k=j+1}^{\infty} \frac{D^{2k-1}}{\Gamma(k+\frac{1}{2})} = \frac{(2j-1)!!}{2^j} e^{-D^2} \sum_{k=j}^{\infty} \frac{D^{2k+1} 2^{k+1}}{(2k+1)!!}. \tag{51}$$

Inserting in (48) gives

$$I(\delta) = A_0 \sum_{j=0}^{\infty} \frac{\Psi^{(2j)}(\frac{1}{2})}{(2v)^{2j+1}(2j)!} G(D, j). \tag{52}$$

To obtain the derivatives of $\Psi(\tau)$, we define $\psi(\tau) = \log \Psi(\tau)$. The first two derivatives of $\varphi(\tau)$ are contained in $\upsilon(\tau) = \log \Upsilon(\tau)$. Therefore,

$\psi(\frac{1}{2}) = \psi'(\frac{1}{2}) = \psi''(\frac{1}{2}) = 0$, $\psi^{(k)}(\frac{1}{2}) = \varphi^{(k)}(\frac{1}{2})$ for $k > 2$, and

$$\Psi^{(k)}(\tfrac{1}{2}) = \varphi^{(k)}(\tfrac{1}{2}) + \sum_{j=3}^{k-3} \binom{k-1}{j-1} \Psi^{(k-j)}(\tfrac{1}{2}) \varphi^{(j)}(\tfrac{1}{2}), \tag{53}$$

where it is understood that the summation is zero when the upper limit is less than the lower limit. The value of D to insert in (52) is $2v\delta$, which can be simplified when $\delta <$ ½ to

$$D = 2v\delta = \operatorname{erfc}^{-1}(\varepsilon). \tag{54}$$

The fact that D depends only on the desired precision ε, makes it advantageous to store pre-calculated values of G(D, j) in a table for suitable values of ε.

The third choice, $\Upsilon(\tau) = \Upsilon_3(\tau)$, gives

$$I(\delta) = 2A_0 \sum_{j=0}^{\infty} \frac{\Psi^{(2j)}(\frac{1}{2})}{(2j)!} \int_0^{\frac{1}{2}+\delta} \{4\tau(1-\tau)\}^b \left(\tau - \tfrac{1}{2}\right)^{2j} d\tau. \tag{55}$$

Substituting $\theta = 4(\tau - \frac{1}{2})^2$ gives

$$I(\delta) = \frac{A_0}{2}\sum_{j=0}^{\infty}\frac{\Psi^{(2j)}(\frac{1}{2})}{4^j(2j)!}\int_0^{4\delta^2}(1-\theta)^b\theta^{j-\frac{1}{2}}d\theta = \frac{A_0}{2}\sum_{j=0}^{\infty}\frac{\Psi^{(2j)}(\frac{1}{2})}{4^j(2j)!}\mathrm{B}_{4\delta^2}(j+\tfrac{1}{2}, b+1), \tag{56}$$

where B_y denotes the incomplete Beta function

$$\mathrm{B}_y(\alpha,\beta) = \int_0^y t^{\alpha-1}(1-t)^{\beta-1}dt\,. \tag{57}$$

To obtain the derivatives of $\Psi(\tau)$, we note that the derivatives of $\upsilon(\tau) = \log \Upsilon(\tau)$ can be calculated in the same way as the derivatives of $\varphi(\tau)$. Therefore, the derivatives of $\psi(\tau) = \varphi(\tau) - \upsilon(\tau)$ can be calculated by adding an extra term to the *i*-sum in (44). For k > 1,

$$\psi^{(k)}(\tau) = \frac{1-k}{\tau}\psi^{(k-1)}(\tau) - \sum_{i=1}^{c+1}\sum_{j=1}^{k}\frac{x_i\eta_{ijk}\tau^{jr\omega_i-k}}{(1-\tau^{r\omega_i})^j}\,, \tag{58}$$

where we define $x_{c+1} = -b$ and $\omega_{c+1} = 1/r$. Now

$$\Psi^{(k)}(\tfrac{1}{2}) = \psi^{(k)}(\tfrac{1}{2}) + \sum_{j=3}^{k-3}\binom{k-1}{j-1}\Psi^{(k-j)}(\tfrac{1}{2})\psi^{(j)}(\tfrac{1}{2})\,,\ \Psi(½)=1,\ \Psi'(½)=0,\ \Psi''(½)=0. \tag{59}$$

An experimental evaluation of the three Taylor methods described here finds that the second and third methods have considerably faster convergence than the first method. The convergence may be satisfactory even when δ = ½. In many cases, the third method has the fastest convergence. Nevertheless, the second method may be preferred for economic reasons since it can use pre-calculated values of $\mathrm{G}(D, j)$ when δ < ½, while the third method requires the more time-consuming calculations of the incomplete Beta function. The third method is preferred when δ = ½ where the incomplete Beta function is replaced by the complete Beta function.

To summarize, $\mathrm{mwnchypg}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega})$ can be calculated with approximate precision ε, if δ < ½, by using equations (36), (37), (46), (43) or (44), (53), (54), (51), (52), (3) and (2). It is recommended to scale the sum terms in (52) with $A_0\Lambda(\mathbf{x})$ to avoid numeric overflow and underflow. The evaluations of the equations (36), (37), (43) and (44) all involve expressions of the type $(1-2^y)^j$. Appropriate Taylor expansions are recommended to avoid loss of precision in these expressions when *y* is near zero or large negative.

The convergence may be poor when *d* is low. As an aid for predicting how good the convergence is, we define the normalized reciprocal *d*:

$$E = \frac{\boldsymbol{\omega}\cdot\mathbf{m}}{\boldsymbol{\omega}\cdot(\mathbf{m}-\mathbf{x})} = \frac{1}{d}\sum_{i=1}^{c}\omega_i m_i\,. \tag{60}$$

A high value of *E* indicates a distorted integrand curve (*E* = 5 for fig. 2a, and 14986 for fig 2b). It has been found experimentally that the convergence of the second and third method is good when δ < 0.25 and *E* < 10.

As a corollary to the expansion formulas, we may obtain reasonable approximations by truncating the series. Truncating the expansion (56) to its first term with δ = ½ gives the

approximation

$$\text{mwnchypg}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) \approx \Lambda(\mathbf{x}) A_0 \frac{\sqrt{\pi}\,\Gamma(b+1)}{2\Gamma(b+\frac{3}{2})}. \tag{61}$$

Truncating the expansion (47) to its first term with $\delta = \infty$ gives the approximation

$$\text{mwnchypg}(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) \approx \Lambda(\mathbf{x}) A_0 \sqrt{\frac{\pi}{-a_2}}. \tag{62}$$

The exponent in the first term of (47) is $\upsilon(\tau)$ which is a 2'nd order Taylor approximation to $\varphi(\tau) = \log \Phi(\tau)$. This approximation method is known as Laplace's method, resulting in (62). According to Bender and Orszag (1978:272), the accuracy of Laplace's method can be improved by adding more terms to the Taylor expansion of $\varphi(\tau)$. The calculation according to Bender and Orszag's method involves nested Taylor sums and possible convergence problems. These disadvantages are avoided here by expanding $\Psi(\tau)$ rather than $\varphi(\tau)$. This method is an improvement to Laplace's method with general applicability to integrals of unimodal functions.

5.4 Continued fraction expansion

The convergence of the abovementioned three expansions can be accelerated considerably by conversion of the Taylor expansions to the corresponding continued fraction expansions in cases where δ is near or equal to ½ and the convergence is poor. No improvement is obtained in cases where the convergence of the Taylor expansion is already good. The continued fraction expansion corresponding to a Taylor expansion is calculated by the method described by Perron (1913). While the Taylor expansion is theoretically convergent for $\delta < ½$, the corresponding continued fraction expansion is theoretically divergent, but practically applicable. The continued fraction method has the disadvantages that it requires the calculation of large Hankel determinants and that it is difficult to evaluate the precision obtained. This method will therefore not be described in further detail here.

5.5 Numerical integration

Numerical integration is needed in cases where none of the abovementioned calculation methods are applicable. The integrand in (4) is not suited for numerical integration because it has most of its weight near 0 where, in most cases, it is not analytic. We prefer to integrate $\Phi(\tau)$ given by (33), using the value of r obtained from (36) and (37). We may improve the performance by integrating $\Phi(\tau)+\Phi(1-\tau)$ over half the interval to take advantage of the fact that $\Phi(\tau)$ is almost symmetric. A Gauss-Legendre method (Evans, 1995) with 4 - 10 points and a variable step length is suitable. The step length should be small where the integrand curve is steepest, which is the endpoints or inflection points, as seen in figure 2a and b.

6. Software implementation

A C++ implementation of the methods described here can be downloaded from www.agner.org/random. An implementation for the R language is available as the package named BiasedUrn from Any CRAN mirror.

7. Suggestions for future research

Better approximations to the mean and variance would be useful. Numerical integration is the only method that is suitable for the cases where *n* and *E* are both high. A more efficient calculation method covering such cases is needed. Equation (46) gives an approximate upper limit to the relative error of the Taylor expansion methods. This error estimate appears to be reliable in practice, but an exact upper limit to the error may be preferable. Fisher's noncentral hypergeometric distribution is more well-researched than Wallenius' distribution. More research on the behavior of the latter is needed.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables.* New York: Dower.

Bender, C. M. and Orszag, S. A (1978). *Advanced Mathematical Methods for Scientists and Engineers.* New York: McGraw-Hill.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research. Vol. 1 - The analysis of case-control studies.* Lyon: International Agency for Research on Cancer.

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.

Evans, G. A. (1995). *Practical Numerical Analysis.* Chichester: John Wiley & Sons.

Fisher, R. A. (1935). The Logic of Inductive Inference. *Journal of the Royal Statistical Society* **98**, 39-82.

Fog, A. (2007). Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions. Working paper. www.agner.org/random/theory/nchyp2.pdf.

Gart, J. J. (1962). Approximate Confidence Limits for the Relative Risk. *Journal of the Royal Statistical Society. Series B (Methodological)* **24**, 454-463.

Gart. J. J. (1987). The equivalence of two corrections to the approximate mean of an entry in a contingency table. *Biometrika* **74**, 661-663.

Graves, T. and Hamada, M. (2006). Biased Reduced Sampling: Detectability of an Attribute and Estimation of Prevalence. *Quality and Reliability Engineering International* **2**, 385-392.

Hannan, J. and Harkness, W. (1963). Normal Approximation to the Distribution of Two Independent Binomials, Conditional on Fixed Sum. *Annals of Mathematical Statistics* **34**, 1593-1595.

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

Hernández-Suárez, C. M. and Castillo-Chavez, C. (2000). Urn models and vaccine efficacy estimation. *Statistics in Medicine* **19**, 827-835.

Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics, 1: Discrete Distributions.* Boston: Houghton Mifflin Co.

Johnson, N. L.; Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions, Second Edition.* New York: Wiley-Interscience.

Johnson, N. L.; Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* New York: Wiley-Interscience.

Levin, B. (1984). Simple Improvements on Cornfield's Approximation to the Mean of a Noncentral Hypergeometric Random Variable. *Biometrika* **71**, 630-632.

Lyons, N. I. (1980). Closed Expressions for Noncentral Hypergeometric Probabilities. *Communications in Statistics, B: Simulation and Computation* **9**, 313-314.

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

Manly, B. F. J. (1985). The Statistics of Natural Selection on Animal Populations. London: Chapman and Hall.
Manly, B. F. J., Miller, P., and Cook, L. M. (1972). Analysis of a Selective Predation Experiment. *The American Naturalist* **106** (952) 719-736.
Marriott, F. C. H. (1990). *A dictionary of statistical terms.* 5th edition. New York: Longman.
McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. London: Chapman & Hall.
Perron, O. (1913). *Die Lehre von den Kettenbrüchen*. Leipzig: B. G. Teubner.
SAS Institute (2002). *SAS OnlineDoc*, 9th ed. Cary, NC: SAS Institute Inc.
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.