doi:10.3969/j.issn.1008-1399.2024.03.019

# 计算负超几何概率分布的迭代算法

王福昌
(防灾科技学院 基础部, 河北 三河 065201)

摘 要 当负超几何分布的中参数变大时,为克服计算其概率分布律时需要计算大阶乘而产生的计算机上溢问题,给出计算概率分布律的简便迭代公式和计算机程序,通过数值分析和比较,验证了递推公式和程序的正确性.

关键词 超几何分布; 负超几何分布; 概率分布律
中图分类号 O171 文献标识码 A 文章编号 1008-1399(2024)03-0061-03

# An Iterative Algorithm of Calculating Probability Mass Function for Negative Hypergeometric Distribution

WANG Fuchang
(Department of Basic Courses, Institute of Disaster Prevention, Sanhe 065201, China)

**Abstract** To address the issue of computer overflow resulting from the evaluation of large factorials, this paper presents a straightforward iterative formula and computer program for calculating the probability mass function of the negative hypergeometric distribution. The correctness of both the formula and program is validated through numerical analysis and comparison.

**Keywords** hypergeometric distribution, negative hypergeometric distribution, probability distribution

## 1 引言

负超几何分布是一种离散型概率分布,它描述的是为了获得固定数目的成功次数而要进行的试验次数的概率分布. Guenther W (1975)[1]定义了负超几何分布,并研究了负超几何分布和超几何分布间的联系,描述了负超几何分布在实际中的应用. Piccolo D(2002)[2]对负超几何分布参数的最大似然估计的渐近方差做了估计. D'Elia A(2004)等人[3]给出了负超几何分布的矩估计量.

负超几何分布可以用于数据分析[3]、现代密码学[4,5]和工农业生产中,但因为涉及阶乘和组合运算,其概率分布的计算非常复杂,特别是当试验次数较大时,直接计算会导致上溢. 现有的统计软件,如 SAS Enterprise Guide、Minitab、SPSS、MATLAB 和 Excel 等,都没有计算负超几何概率值的功能[6]. López-Blázquez (2001)[7]给出了负超几何分布和负二项式分布间的近似关系, Teerapabolarn K (2011)[8]给出了负超几何分布的一种 Poisson 近似,然而负超几何分布的负二项分布和 Poisson 分布近似[9]都需要参数满足一定的使用条件.

为了解决负超几何分布的计算问题,避免试验次数较大时,计算大阶乘出现的计算机上溢问题,常采用如下两种解决方案. 一是阶乘对数法:利用阶乘对数表或对数函数求负超几何分布概率分布律的对数,然后利用指数函数求其反对数. 显然,直接采用这种方法当阶乘数很大时计算量增大,解决方案是利用斯特林阶乘近似公式 $n!\approx\sqrt{2\pi}n^{n+\frac{1}{2}}\mathrm{e}^{-n}$ 计算. 彭求实(1997)[10]为提高计算的精度,将斯特林公式改进为 $n!\approx\sqrt{2\pi}n^{n+\frac{1}{2}}\mathrm{e}^{-n+\frac{1}{12n}-\frac{1}{360n^3}}$. 这种算法的计算误差中算法的截断误差和计算机浮点运算的舍入

收稿日期: 2022-08-23 修改日期: 2023-02-06
基金项目: 廊坊市科技局科学研究与发展计划项目(2022011031),防灾科技学院教育研究与教学改革项目(JY2022A03).
作者简介: 王福昌(1974-),男,山东菏泽人,硕士,教授,研究方向:概率统计,Email:fzmath@126.com.

误差. 二是递推公式法:利用负超几何分布的性质,找到递推公式. 这种算法不会出现计算机上溢,因而计算结果更加精确.

本文首先介绍了负超几何分布的两种不同的概率分布律公式,证明了它们的等价性. 然后基于概率分布公式,利用组合数性质,给出计算的递推公式. 最后给出了计算的程序代码,通过与文献结果比较,验证了算法的正确性和有效性.

## 2 预备知识

首先给出负超几何分布两种定义等价性的严格证明.

**定义1** 从一个装有 $m$ 个白球,$n$ 个黑球的袋子中不放回地摸球,以 $X$ 表示摸到 $r$ 个白球时所需的摸球次数,$X$ 的概率分布为

$$P(X=k)=\frac{C_{k-1}^{r-1}C_{m+n-k}^{m-r}}{C_{m+n}^{m}},\ k=r,r+1,\cdots,r+n, \tag{1}$$

则称 $X$ 服从负超几何分布,记为 $X\sim NH(m,n,r)$.

负超几何分布也可以看作是一种离散的等待时间分布,即等待第 $r$ 个白球出现时的试验(抽取)次数. 下面给出另一种形式的定义。

**定义2** 从一个装有 $m$ 个白球,$n$ 个黑球的袋子中依次不放回地摸球,以 $X$ 表示摸到 $r$ 个白球时所需的摸球次数,$X$ 的概率分布为

$$P(X=k)=\frac{C_{m}^{r-1}C_{n}^{k-r}}{C_{m+n}^{k-1}}\cdot\frac{m-r+1}{m+n-k+1},\quad k=r,r+1,\cdots,r+n, \tag{2}$$

则称 $X$ 服从负超几何分布,记为 $X\sim NH(m,n,r)$.

若令 $N=m+n$,$p=\lim\limits_{N\to\infty}m/N$,则负超几何分布近似为负二项分布(也称帕斯卡分布).

如果是有放回地摸球,则等待第 $r$ 个白球出现时的试验(抽取)次数 $X$ 服从负二项分布,有

$$P(X=k)=C_{k-1}^{r-1}p^{r}(1-p)^{k-r},\ k=r,r+1,\cdots \tag{3}$$

## 3 主要结果和算法

### 3.1 两种定义的等价性

首先给出负超几何分布两种定义等价性的严格证明.

**定理1** 若随机变量 $X\sim NH(m,n,r)$,则有

$$\begin{aligned}P(X=k)&=\frac{C_{m}^{r-1}C_{n}^{k-r}}{C_{m+n}^{k-1}}\cdot\frac{m-r+1}{m+n-k+1}\\&=\frac{C_{k-1}^{r-1}C_{m+n-k}^{m-r}}{C_{m+n}^{m}},\ k=d,d+1,\cdots,r,\end{aligned} \tag{4}$$

**证明** 利用阶乘和组合数性质,有

$$\begin{aligned}P\{X=k\}&=\frac{C_{m}^{r-1}C_{n}^{k-r}}{C_{m+n}^{k-1}}\cdot\frac{m-r+1}{m+n-k+1}\\&=\frac{m!}{(r-1)!(m-r+1)!}\frac{n!}{(k-r)!(n-k+r)!}\\&\quad\cdot\frac{(k-1)!(m+n-k+1)!}{(m+n)!}\cdot\frac{m-r+1}{m+n-k+1}\\&=\frac{m!}{(r-1)!(m-r)!}\cdot\frac{n!}{(x-r)!(n-k+r)!}\\&\quad\cdot\frac{(k-1)!(m+n-k)!}{(m+n)!}\\&=\frac{m!}{(r-1)!(m-r)!}\cdot\frac{n!}{(k-r)!(n-k+r)!}\\&\quad\cdot\frac{(k-1)!(m+n-k)!}{(m+n)!}\\&=\frac{C_{k-1}^{r-1}C_{m+n-k}^{m-r}}{C_{m+n}^{m}},\ k=r,r+1,\cdots,r+n.\end{aligned}$$

在一定条件下,负超几何分布可以用负二项分布或 Poisson 分布近似[5,7-9],然而近似计算中也会面临大阶乘的计算难题,而且精度和参数的适用范围都需要考虑,这里利用超几何分布的性质给出一种直接的直接迭代方法,可以获得高精度的概率分布.

### 3.2 求负超几何分布的递推算法

为便于推导,按照定义1中负超几何分布的公式进行讨论.

**定理2** 设随机变量 $X\sim NH(m,n,r)$,若记

$$NH(k;m,n,r)=P(X=k)=\frac{C_{k-1}^{r-1}C_{m+n-k}^{m-r}}{C_{m+n}^{m}},\quad k=r,r+1,\cdots,r+n,$$

则有递推公式为

$$NH(k;m,n,r)=NH(k-1;m,n,r)\cdot\frac{(k-1)(n-k+r+1)}{(k-r)(m+n-k+1)},\ k=r+1,\cdots,r+n. \tag{5}$$

**证明** 利用负超几何分布的概率分布表达式和组合数性质,可得

$$\begin{aligned}\frac{NH(k;m,n,r)}{NH(k-1;m,n,r)}&=\frac{\frac{C_{k-1}^{r-1}C_{m+n-k}^{m-r}}{C_{m+n}^{m}}}{\frac{C_{k-2}^{r-1}C_{m+n-k+1}^{m-r}}{C_{m+n}^{m}}}=\frac{C_{k-1}^{r-1}}{C_{k-2}^{r-1}}\frac{C_{m+n-k}^{m-r}}{C_{m+n-k+1}^{m-r}}\\&=\frac{\frac{(k-1)!}{(r-1)!(k-r)!}}{\frac{(k-2)!}{(r-1)!(k-r-1)!}}\cdot\frac{\frac{(m+n-k)!}{(m-r)!(n-k+r)!}}{\frac{(m+n-k+1)!}{(m-r)!(n-k+1+r)!}}\\&=\frac{(k-1)}{(k-r)}\cdot\frac{(n-k+r+1)}{(m+n-k+1)}\end{aligned}$$

故递推公式为

$$NH(k;m,n,r)=NH(k-1;m,n,r)\cdot\frac{(k-1)}{(k-r)}\cdot\frac{(n-k+r+1)}{(m+n-k+1)},\ k=r+1,\cdots,r+n.$$

有了上面的递推公式(5)，使用时还需要计算初值，即 $k=r$ 时，计算

$$NH(r;m,n,r)=\frac{C_{m+n-r}^{m-r}}{C_{m+n}^{m}}.$$

下面给出初值计算的迭代公式.

**定理 3**　设随机变量 $X\sim NH(m,n,r)$，若 $k=r$，则 $NH(r;m,n,r)=\frac{C_{m+n-r}^{m-r}}{C_{m+n}^{m}}$，计算 $NH(r;m,n,r)$ 的递推公式为

$$\begin{cases}NH(1;m,n,r)=\dfrac{n}{m+n}\\ NH(j;m,n,r)=NH(j-1;m,n,r)\cdot\dfrac{n-j+1}{m+n-j+1},\ j=2,3,\cdots,r.\end{cases}\tag{6}$$

**证明**　因为

$$NH(r;m,n,r)=\frac{C_{m+n-r}^{m-r}}{C_{m+n}^{m}}=\frac{(m+n-r)!}{(m-r)!\,n!}\cdot\frac{m!\,n!}{(m+m)!}$$

$$=\frac{m}{(m+n)}\cdot\frac{(m-1)}{(m+n-1)}\cdot\cdots\cdot\frac{(m-r+1)}{(m+n-r+1)}$$

使用计算 $NH(r;m,n,r)$ 的递推公式为

$$NH(j;m,n,r)=NH(j-1;m,n,r)\cdot\frac{n-j+1}{m+n-j+1},\ j=2,3,\cdots,r.$$

因为概率值一定不超过 1，所以使用递推算法不存在数值的计算机上溢问题. 这种方法从理论上说是一种精确算法，计算误差只可能是来自乘法的舍入误差.

### 3.3　求负超几何分布的递推算法

下面给出相应的算法如下：

**Step 1**　给定初始参数 $m,n,r$，用公式(6)计算初始值 $NH(r;m,n,r)$；

**Step 2**　对 $k=r+1,\cdots,r+n$，利用初值 $NH(r;m,n,r)$ 和递推公式(5)，计算负超几何分布的概率 $NH(k;m,n,r)$.

### 3.4　迭代算法的 MATLAB 程序

为便于计算，编写计算负超几何分布的概率分布的 MATLAB 函数 nehypgeopmf，使用时直接调用即可.

```
function pmfk = nehypgeopmf(k,m,n,r)
% nehypgeopmf(k,m,n,r) returns thenegative hypergeometric probability
%    mass function at k with integer parametersm,n,r.
%%计算初值 nehypgeopmf(r,m,n,r)
NHrmnr = m/(m+n);
for j = 1:r-1
    NHrmnr = NHrmnr*(m-j)/(m+n-j);
end
pmfk = NHrmnr;
%迭代公式 nehypgeopmf(k,m,n,r) =
%nehypgeopmf(k-1,m,n,r)*(k-1)/(k-r)*(n-k+r+1)/(m+n-k+1)
if (k>=r)&&(k<=r+n)
for i = r+1:k
        pmfk =  pmfk*(i-1)/(i-r)*(n-i+r+1)/(m+n-i+1);%迭代公式
    end
else
    pmfk = 0;
end
```

### 3.5　数值计算结果和比较

为了验证算法和程序的正确性，首先在 $m,n,r$ 较小的时候，将程序计算的结果与按公式(3)和(4)计算的结果进行比较，发现计算结果一致，说明算法和程序是正确的. 为进一步验证，又与文献[5]的结果进行对比.

**表 1　结果比较**

| 参数 | $k=2,m=10,n=10,r=1$ | $k=6,m=10,n=90,r=1$ |
|---|---|---|
| 文献结果 | 0.2631578947368 | 0.0614476175710 |
| 本文结果 | 0.263157894736842 | 0.061447617571174 |
| 参数 | $k=100,m=1000,n=1000,r=50$ | $N=10000,M=2000,u=100,r=50$ |
| 文献结果 | 0.0408281487021 | 6.1609683854030e−12 |
| 本文结果 | 0.040828148702016 | 6.160968385403023e−12 |

文献[5]中给的结果是小数点后 13 位，本文是 15 位，可以看出第 1 组和中小数点后有 13 位数字完全相同，第 2 组和第 3 组参数中的小数点后第 13 位数字不同，第 4 组参数中本文结果也与文献结果一致。可见本文的算法结果正确，精度很高.

## 4　结果与讨论

负超几何分布概率分布快速高精度计算是一项基础的工作，为编制任意参数下的概率分布表、生成

(下转第 74 页)

数，其中 $P(z)=c_0z^m+c_1z^{m-1}+\cdots+c_m(c_0\neq0)$ 与 $Q(z)=b_0z^n+b_1z^{n-1}+\cdots+b_n(b_0\neq0)$ 为互质的多式，且满足条件：(1) $n-m\geqslant2$；(2)在实轴上，$Q(z)\neq0$. $a,\bar{a}$ 分别为 $R(z)$ 在上半平面内与下半平面内唯一的孤立奇点(即极点)，则

(i)若 $R(\bar{z})=\overline{R(z)}$，有

$\mathrm{Re}\{\mathrm{Res}[R(z),a]\}=\mathrm{Re}\{\mathrm{Res}[R(z),\bar{a}]\}=0$；

$\mathrm{Im}\{\mathrm{Res}[R(z),a]\}=-\mathrm{Im}\{\mathrm{Res}[R(z),\bar{a}]\}$.

(ii)若 $R(\bar{z})=-\overline{R(z)}$，有

$\mathrm{Im}\{\mathrm{Res}[R(z),a]\}=\mathrm{Im}\{\mathrm{Res}[R(z),\bar{a}]\}=0$；

$\mathrm{Re}\{\mathrm{Res}[R(z),a]\}=-\mathrm{Re}\{\mathrm{Res}[R(z),\bar{a}]\}$.

如例2满足推论2(ii)条件，故有

$$\mathrm{Im}\{\mathrm{Res}[R(z),\mathrm{i}]\}=\mathrm{Im}\{\mathrm{Res}[R(z),-\mathrm{i}]\}=0;$$

$$\mathrm{Re}\{\mathrm{Res}[R(z),\mathrm{i}]\}=-\mathrm{Re}\{\mathrm{Res}[R(z),-\mathrm{i}]\}\left(=\frac{1}{4}\right).$$

**注2** 若定理3及推论2中的条件“在实轴上，$Q(z)\neq0$”不成立，则结论未必成立，反例如下.

设函数 $R(z)=\dfrac{P(z)}{Q(z)}=\dfrac{1}{(z^2+1)^2\mathrm{i}z}$，则 $R(z)$ 在复平面上除两个共轭二级极点 $z=\pm\mathrm{i}$ 外，还有一个一级极点 $z=0$(即 $Q(z)$ 有实零点)，且 $R(\bar{z})=-\overline{R(z)}=\dfrac{1}{(\bar{z}^2+1)^2\mathrm{i}\bar{z}}$，由引理1，得

$$\begin{aligned}\mathrm{Res}[R(z),\mathrm{i}]&=\lim_{z\to\mathrm{i}}\left[(z-\mathrm{i})^2\,\frac{1}{(z^2+1)^2\mathrm{i}z}\right]'\\&=\lim_{z\to\mathrm{i}}\left[\frac{1}{(z+\mathrm{i})^2\mathrm{i}z}\right]'\\&=-\lim_{z\to\mathrm{i}}\frac{2(z+\mathrm{i})\mathrm{i}z+\mathrm{i}(z+\mathrm{i})^2}{[(z+\mathrm{i})^2\mathrm{i}z]^2}=\frac{\mathrm{i}}{2}.\end{aligned}$$

由定理1(ii)得

$$\mathrm{Res}[R(z),-\mathrm{i}]=-\overline{\mathrm{Res}[R(z),\mathrm{i}]}=\frac{\mathrm{i}}{2}.$$

于是

$$\mathrm{Im}\{\mathrm{Res}[R(z),\mathrm{i}]\}=\mathrm{Im}\{\mathrm{Res}[R(z),-\mathrm{i}]\}=\frac{1}{2}\neq0.$$

## 参考文献


[1] 钟玉泉. 复变函数论[M]. 4版. 北京：高等教育出版社，2013：167－168.

[2] 路见可，钟寿国，刘士强. 复变函数[M]. 2版. 武汉：武汉大学出版社，2018.

[3] 李明泉. 实系数有理分式函数的共轭复极点的留数[J]. 安庆师范学院学报(自然科学版)，2007，13(4)：96－98.

[4] 余家荣. 复变函数[M]. 5版. 北京：高等教育出版社，2014.

[5] 王镇英. 留数在无穷积分中的应用[J]. 工科数学，1993(S2)：217－218.

[6] 王文帅. 关于用留数定理计算实积分问题的教学研究[J]. 大学教育，2019(4)：106－108.

[7] 刘云冰. 用留数计算 $\int_{-\infty}^{+\infty}R(x)\mathrm{e}^{\mathrm{i}ax}\mathrm{d}x$ 型积分注记[J]. 高等数学研究，2007(01)：99－100.


(上接第63页)

服从负超几何分布的随机数、参数估计和进一步广泛应用打下了基础. 本文给出的递推算法，计算机实现简单，计算速度快，精度高，在参数取很大值时，避免计算大阶乘，能很快得到高精度概率分布.

## 参考文献


[1] Guenther W. The Inverse Hypergeometric-a useful model [J]. Statistica Neerlandica, 2010, 29(4): 129－144.

[2] Piccolo D. Some Approximations for the Asymptotic Variance of the Maximum Likelihood Estimator of the Parameter in the Inverse Hypergeometric Random Variable[J]. 2001, (3): 199－213.

[3] D'Elia A, Piccolo D. A Mixture Model for Preferences Data Analysis[J]. Computational Statistics & Data Analysis, 2005, 49(3): 917－934.

[4] 胡冬萍. 基于负超几何分布的十进制分组加密方案研究[D]. 武汉：华中科技大学，2015.

[5] Hu D P. An Improved Negative Binomial Approximation with High Accuracy to the Negative Hypergeometric Probability for Order-Preserving Encryption. Journal of Difference Equations and Applications, 2017, 23(2): 88－99.

[6] Miller G K, Fridell S L. A Forgotten Discrete Distribution? Reviving the Negative Hypergeometric Model[J]. The American Statistician, 2007, 61(4): 347－350.

[7] López-Blázquez F, Salamanca-Mio B. Exact and approximated relations between negative hypergeometric and negative binomial probabilities[J]. Communications in Statistics, 2001, 30(5): 957－967.

[8] Teerapabolarn K. On the Poisson Approximation to the Negative Hypergeometric Distribution[J]. Bulletin of the Malaysian Mathematical Sciences Society, 2011, 34(2): 331－336.

[9] 戴朝寿，候艳艳. 负超几何分布、负二项分布与 Poisson 分布之间的关系及其推广[J]. 南京大学学报：数学半年刊，2001，18(1)：76－84.

[10] 彭求实. 两类有关产品检测古典概率的近似计算[J]. 数理统计与管理，1997，16(4)：33－36.