超几何分布、二项分布与正态分布的区别与联系

◎郭　婧[1]　李　强[2]（1.青岛西海岸新区第一高级中学，山东　青岛　266000；
2.青岛大学，山东　青岛　266000）

【摘要】人教A版选修2-3中介绍了超几何分布、二项分布和正态分布，前两者属于离散型随机变量服从的分布，后者属于连续型随机变量服从的分布.实际中的许多问题都可以利用这三个概率模型来解决.区分前两者的关键是看属于“不放回”模型还是“有放回”模型.同时，随着产品数量的增加，超几何分布越来越趋近于二项分布；随着试验次数的增加，二项分布越来越趋近于正态分布.从而三者在极限方面实现统一.

【关键词】超几何分布；二项分布；正态分布；极限

【基金项目】山东省教育学会科技教育专项课题：基于虚拟现实的高中数学翻转课堂教学模式研究（课题号18-KJJY-0074）.科技部国家重点研发计划：流域水系分级嵌套耦合大规模水文模拟并行算法设计（No.2017YFB0203102）.

一、总述

人教A版选修2-3中介绍了超几何分布、二项分布和正态分布，前两者属于离散型随机变量服从的分布，后者属于连续型随机变量服从的分布.在实际教学中发现学生辨别这些分布是难点，或者即使能辨别却无法从本质上认识它们.本文介绍三种分布的区别与联系，来帮助学生克服此难点.

在教材中，三种分布的定义如下：

一般地，在含有M件次品的N件产品中，任取n件，其中恰有X件次品，则

$$P(X=k)=\frac{C_M^kC_{N-M}^{n-k}}{C_N^n},k=0,1,2\cdots,m,$$

其中$m=\min\{M,n\}$，$n\leqslant N,M\leqslant N,n,M,N\in\mathbf{N}^*$.如果随机变量$X$的分布列具有上述形式，则称随机变量$X$服从超几何分布（hypergeometric distribution）.

一般地，在n次独立重复试验中，用X表示事件A发生的次数，设每次试验中事件A发生的概率为p，则$P(X=k)=C_n^kp^k(1-p)n-k,k=0,1,2,\cdots n$.此时称随机变量$X$服从二项分布（binomial distribution），记作$X:B(N,p)$，并称p为成功概率.

一般地，如果对于任何实数$a,b(a<b)$，随机变量X满足

$P(a<X\leqslant b)=\int_a^b\varphi_{\mu,\sigma}(x)\mathrm{d}x$，其中$\varphi_{\mu,\sigma}(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},x\in(-\infty,+\infty)$，则称随机变量$X$服从正态分布（normal distribution），记作$X\sim N(\mu,\sigma^2)$[1].

二、超几何分布与二项分布的区别与联系

超几何分布是“不放回”情境中的古典概型，二项分布是“有放回”情境中的n次独立重复试验概型.如教材习题2.2B组第三题：某批n件产品的次品率为2%，现从中任意地抽出3件进行检验，问：当$n=500$，5 000，50 000时，分别以放回和不放回的方式抽取，恰好抽到1件次品的概率各是多少？设抽取到的次品数为X.若以有放回的方式抽取，抽取3次相当于做了3次独立重复试验，则$X:B(3,2\%)$，故$P(X=1)=C_3^1\times2\%\times(1-2\%)^2\approx0.057\ 624$.若以不放回方式抽取，这个问题回归到古典概型，$X$服从超几何分布.$n=500$时，次品数是$500\times2\%=10$，$P(X=1)=\frac{C_{10}^1C_{490}^2}{C_{500}^3}\approx0.057\ 853$.同理可得$n=5\ 000$时，$P(X=1)=\frac{C_{100}^1C_{4900}^2}{C_{5000}^3}\approx0.057\ 647$.$n=50\ 000$时，$P(X=1)=\frac{C_{1000}^1C_{49000}^2}{C_{50000}^3}\approx0.057\ 626$.由此可见，随着产品数$n$的增加，超几何分布的概率是越来越接近于二项分布的概率的.此结论也可以通过以下表格验证：

n \ k	0	1	2	3
500	-1.16×10^{-4}	2.29×10^{-4}	1.11×10^{-4}	-2.21×10^{-6}
5 000	-1.2×10^{-5}	2.3×10^{-5}	1.11×10^{-5}	-2.34×10^{-7}
50 000	-1.1×10^{-6}	2×10^{-6}	1.11×10^{-6}	-2.35×10^{-8}

其中表中的每一个数表示相应超几何分布与二项分布的概率差.由此可见,无论取几件次品,随着产品数的增加,二者的差距都是越来越趋近于零的[2].即

$$\lim_{N\to\infty}\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}=C_n^k p^k(1-p)n-k.$$

下面我们对这个结论给出证明:

$$\lim_{N\to\infty}\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$$

$$=\lim_{N\to\infty}\frac{A_M^k A_{N-M}^{n-k}n!}{k!\ (n-k)!\ A_N^n}$$

$$=\lim_{N\to\infty}C_n^k\frac{M(M-1)\cdots(M-k+1)(N-M)(N-M-1)(N-M-n+k+1)}{N(N-1)\cdots(N-n+1)}$$

$$=\lim_{N\to\infty}C_n^k\frac{pN(pN-1)\cdots(pN-k+1)(N-pN)(N-pN-1)(N-pN-n+k+1)}{N(N-1)\cdots(N-n+1)}$$

$$=\lim_{N\to\infty}C_n^k\frac{\overbrace{pN(pN-1)\cdots(pN-k+1)}^{k\text{个}}}{N^k}\cdot$$

$$\frac{\overbrace{(N-pN)(N-pN-1)(N-pN-n+k+1)}^{(n-k)\text{个}}}{N^{n-k}}\cdot$$

$$\frac{N^n}{\underbrace{N(N-1)\cdots(N-n+1)}_{n\text{个}}}$$

$$=C_n^k p^k(1-p)^{n-k}.$$

一般地,如果随机变量依分布收敛于随机变量,则前者期望与方差的极限分别是后者的期望与方差[3].实际上,若X服从超几何分布,则

$$\begin{aligned}E(X)&=\sum_{k=0}^{m}\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}\\&=\sum_{k=0}^{m}\frac{MC_{M-1}^{k-1}C_{N-M}^{n-k}}{NC_{N-1}^{n-1}}\\&=\frac{nM}{N}\sum_{k=0}^{m}\frac{C_{M-1}^{k-1}C_{N-M}^{n-k}}{C_{N-1}^{n-1}}\\&=n\frac{M}{N}\\&=np.\end{aligned}$$

$$\begin{aligned}D(X)&=E(X^2)-(E(X))^2\\&=\sum_{k=0}^{m}k^2\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}-\left(n\frac{M}{N}\right)^2\\&=\sum_{k=1}^{m}k(k-1)\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}+\sum_{k=0}^{m}k\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}-\left(n\frac{M}{N}\right)^2\\&=M(M-1)\sum_{k=2}^{m}\frac{C_{M-2}^{k-2}C_{N-M}^{n-k}}{C_N^n}+n\frac{M}{N}-\left(n\frac{M}{N}\right)^2\\&=n(n-1)\frac{M(M-1)}{N(N-1)}\sum_{k=2}^{m}\frac{C_{M-2}^{k-2}C_{N-M}^{n-k}}{C_{N-2}^{n-2}}+n\frac{M}{N}-\left(n\frac{M}{N}\right)^2\\&=n(n-1)\frac{M(M-1)}{N(N-1)}+n\frac{M}{N}-\left(n\frac{M}{N}\right)^2.\end{aligned}$$

$$\therefore\ \lim_{N\to\infty}D(X)=np(1-p).$$

三、二项分布与正态分布的区别与联系

二项分布是离散型随机变量服从的分布,关注随机变量取某一个值时的概率;正态分布是连续型随机变量服从的分布,关注随机变量在某一范围的概率,存在于生活中的方方面面,如“长度测量的误差”“某一地区同年龄人群的身高、体重、肺活量”“在一定条件下生长的小麦的株高、穗长”等等都服从正态分布.但是貌似毫无关联的二项分布与正态分布,存在以下联系:

定理　在n次独立重复试验中,事件A在每次试验中出现的概率是$p(0<p<1)$,μ_n为n次试验中事件A出现的次数,则

$$\lim_{n\to\infty}P\left(\frac{\mu_n-np}{\sqrt{pq}}<x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

上式表明二项分布收敛于正态分布[4].

四、总结

超几何分布、二项分布与正态分布是三个非常重要的、应用广泛的概率模型.实际中的许多问题都可以利用这三个概率模型来解决.区分前两者的关键是看属于“不放回”模型还是“有放回”模型.同时,随着产品数量的增加,超几何分布越来越趋近于二项分布;随着试验次数的增加,二项分布越来越趋近于正态分布.从而三者在极限方面实现统一.

【参考文献】

[1] 李勇.普通高中课程标准实验教科书数学选修2-3[M].北京:人民教育出版社,2011.

[2] 丁曼.超几何分布与二项分布的联系与区别[J].中国课程辅导,2010(7):115-116.

[3] 魏宗舒.概率论与数理统计教程[M].北京:高等教育出版社,2008.

[4] 汪嘉冈.现代概率论基础[M].上海:复旦大学出版社,2005.