二项分布与超几何分布的区别与联系

■河南省信阳市固始县信合外国语高级中学 胡云兵

在人教版《数学选修2—3》的课本中，第二章《随机变量及其分布列》分别介绍了两种离散型随机变量的概率分布：超几何分布与二项分布。通过实例，让同学们认识模型所刻画的随机变量的共同特点，从而建立新的模型，并能运用两个模型解决一些实际问题。然而部分同学不能准确地辨别要解决的问题是超几何分布还是二项分布，对这两个模型的定义不能很好地理解，一遇到含“取”或“摸”的题型，就认为是超几何分布，不加分析，滥用公式。事实上，超几何分布和二项分布确实有着密切的联系，但也有明显的区别。

课本对于超几何分布的定义是这样的：一般地，若一个随机变量X的分布列为$P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$，其中$k=0,1,2,3,\cdots,l$，$l=\min\{n,M\}$，则称$X$服从超几何分布，记为$X\sim H(n,M,N)$。其概率分布如表1。

表1

X	0	1	2	$\cdots$	l
P	$\frac{C_M^0 C_{N-M}^n}{C_N^n}$	$\frac{C_M^1 C_{N-M}^{n-1}}{C_N^n}$	$\frac{C_M^2 C_{N-M}^{n-2}}{C_N^n}$	$\cdots$	$\frac{C_M^l C_{N-M}^{n-l}}{C_N^n}$

对于二项分布的定义是这样的：若随机变量X的分布列为$P(X=k)=C_n^k p^k(1-p)^{n-k}$，其中$0<p<1$，$k=1,2,\cdots,n$，则称$X$服从参数为$n,p$的二项分布，记为$X\sim B(n,p)$。其概率分布如表2。

表2

X	0	1	2	$\cdots$	n
P	$C_n^0(1-p)^n$	$C_n^1 p^1(1-p)^{n-1}$	$C_n^2 p^2(1-p)^{n-2}$	$\cdots$	$C_n^n p^n(1-p)^0$

超几何分布与二项分布都是取非负整数值的离散分布，表面上看，两种分布的概率求法有截然不同的表达式。

课本中对超几何分布的模型建立是这样的：若有N件产品，其中M件是次品，无放回地任意抽取n件，则其中恰有的次品件数X服从超几何分布。而对二项分布则使用比较容易理解的射击问题来建立模型。若将超几何分布的概率模型改成：若有N件产品，其中M件是次品，有放回的任意抽取n件，则其中恰有的次品件数X服从二项分布。在这里，两种分布的差别就在于“有放回”与“无放回”的差别，超几何分布是不放回地抽取，二项分布是有放回地抽取；超几何分布需要知道总体容量，二项分布不需要知道总体容量，但需要知道“成功率”。

教材上用高尔顿板导出了正态分布，事实上在这个试验中，小球落进各个槽中的分布是一个概率$p=0.5$的二项分布，从一个二项分布导出正态分布，正好揭示了二项分布与正态分布之间的联系。

我们知道二项分布的随机变量是离散的，当n取不同值时，它的分布会有差别，随着n的增加，二项分布越来越接近于正态分布。

题型一 超几何分布

例1 某批产品共有10件，已知从该批产品中任取1件，则取到的是次品的概率为$p=0.2$，现从该批产品中任意抽取3件。

(1)求取出的3件产品中恰好有1件是次品的概率；

(2)求取出的3件产品中次品的件数X的概率分布列与期望。

解析：设该批产品中次品有x件，由已知$\frac{x}{10}=0.2$，解得$x=2$。

(1)设取出的3件产品中次品的件数为X，3件产品中恰好有1件次品的概率为$P(X=1)=\frac{C_2^1C_8^2}{C_{10}^3}=\frac{7}{15}$。

(2)由题意知，X可能为0,1,2，则：

$P(X=0)=\frac{C_8^3}{C_{10}^3}=\frac{7}{15}$，$P(X=1)=\frac{C_2^1C_8^2}{C_{10}^3}$

$=\frac{7}{15}$，$P(X=2)=\frac{C_2^2C_8^1}{C_{10}^3}=\frac{1}{15}$。

X 的分布如表 3。

表 3

X	0	1	2
P	$\frac{7}{15}$	$\frac{7}{15}$	$\frac{1}{15}$

则 $E(X)=0\times\frac{7}{15}+1\times\frac{7}{15}+2\times\frac{1}{15}=\frac{3}{5}$。

点评：超几何分布的特征是：只做了一次试验，且这一次试验是在有限个（个数非常明确、具体）元素中选取若干个元素。典型案例是抽球时每次抽完不放回。

例 2 （2021年北京高三期末）在学期末，为了解学生对食堂用餐满意度情况，某兴趣小组按性别采用分层抽样的方法，从全校学生中抽取容量为 200 的样本进行调查。被抽中的同学分别对食堂进行评分，满分为 100 分。调查结果显示：最低分为 51 分，最高分为 100 分。随后，兴趣小组将男、女生的评分结果按照相同的分组方式分别整理成了频数分布表和频率分布直方图，如表 4 和图 1 所示。

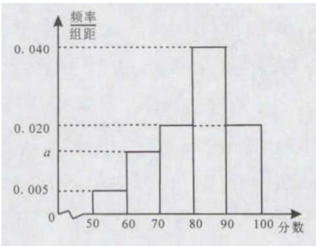

图 1

女生评分结果的频率分布直方图（图 1）。

男生评分结果的频数分布表如表 4。

表 4

分数区间	频数
[50，60)	3
[60，70)	3
[70，80)	16
[80，90)	38
[90，100]	20

为了便于研究，兴趣小组将学生对食堂的评分转换成了“满意度情况”，两者的对应关系如表 5。

表 5

分数	[50，60)	[60，70)	[70，80)	[80，90)	[90，100]
满意度情况	不满意	一般	比较满意	满意	非常满意

（1）求 a 的值；

（2）为进一步改善食堂状况，从评分在 [50，70) 的男生中随机抽取 3 人进行座谈，记这 3 人中对食堂“不满意”的人数为 X，求 X 的分布列。

解析：（1）因为 $(0.005+a+0.020+0.040+0.020)\times10=1$，所以 $a=0.015$。

（2）依题意知，随机变量 X 的所有可能取值为 0，1，2，3。

$$P(X=0)=\frac{C_3^0\cdot C_3^3}{C_6^3}=\frac{1}{20};$$

$$P(X=1)=\frac{C_3^1\cdot C_3^2}{C_6^3}=\frac{9}{20};$$

$$P(X=2)=\frac{C_3^2\cdot C_3^1}{C_6^3}=\frac{9}{20};$$

$$P(X=3)=\frac{C_3^3\cdot C_3^0}{C_6^3}=\frac{1}{20}。$$

所以随机变量 X 的分布列如表 6 所示。

表 6

X	0	1	2	3
P	$\frac{1}{20}$	$\frac{9}{20}$	$\frac{9}{20}$	$\frac{1}{20}$

点评：“在 6 个元素中选取 3 个元素”，符合超几何分布的特征。

题型二 二项分布

例 3 （2021年太原高三5月模拟）人的体重是人的身体素质的重要指标之一。某校抽取了高二的部分学生，测出他们的体重（kg），体重在 40 kg 至 65 kg 之间，按体重进行如下分组：第 1 组 [40，45)，第 2 组 [45，50)，第 3 组 [50，55)，第 4 组 [55，60)，第 5 组 [60，65]，并制成如图 2 所示的频率（视为概率）分布直方图。已知第一组与第三组的频率之比为 1∶3，第 3 组的频数为 90。

（1）求该校抽取的学生总数以及第 2 组的频率；

（2）用这些样本数据估计全市高二学生（学生数众多）的体重，若从全市高二学生中

任选 5 人，设 X 表示这 5 人中体重不低于 55 kg 的人数，求 X 的分布列和数学期望。

解析：(1)设该校抽查的学生总人数为 n，第二组、第三组的频率分别为 p_2、p_3。

则 $p_3=0.025\times3\times5=0.375$，所以 $n=\frac{90}{p_3}=240$。

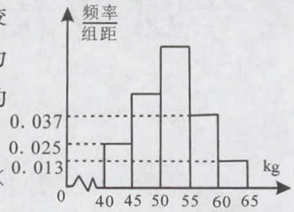

图 2

由 $p_2+0.375+(0.025+0.013+0.037)\times5=1$，解得 $p_2=0.25$。

所以该校抽查的学生总人数为 240 人，从左到右第二组的频率为 0.25。

(2)由(1)知体重不低于 55 kg 的学生的概率为 $p=(0.013+0.037)\times5=\frac{1}{4}$。

X 服从二项分布 $X\sim B\left(5,\frac{1}{4}\right)$，$P(X=k)=C_5^k\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{5-k}$，$k=0,1,2,3,4,5$。

$P(X=0)=C_5^0\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^5=\frac{243}{1\ 024}$；

$P(X=1)=C_5^1\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^4=\frac{405}{1024}$；

$P(X=2)=C_5^2\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^3=\frac{270}{1\ 024}$；

$P(X=3)=C_5^3\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^2=\frac{90}{1\ 024}$；

$P(X=4)=C_5^4\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)=\frac{15}{1\ 024}$；

$P(X=5)=C_5^5\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right)^0=\frac{1}{1\ 024}$。

所以随机变量 X 的分布列如表 7 所示。

表 7

X	0	1	2	3	4	5
P	$\frac{243}{1\ 024}$	$\frac{405}{1\ 024}$	$\frac{270}{1\ 024}$	$\frac{90}{1\ 024}$	$\frac{15}{1\ 024}$	$\frac{1}{1\ 024}$

$E(X)=5\times\frac{1}{4}=\frac{5}{4}$。

点评：当题目当中出现“将频率视为概率”，“某某与某某相互独立”，“从大量(且不知道大量具体的数值)中选取少数几个个体”等关键字眼时，该分布即为二项分布。

例 4 某工厂的某种产品成箱包装，每箱 200 件，每一箱产品在交付用户之前要对产品进行检验，如检验出不合格品，则更换为合格品。检验时，先从这箱产品中任取 20 件检验，再根据检验结果决定是否对余下的所有产品检验。设每件产品为不合格品的概率都为 $p(0<p<1)$，且各件产品是否为不合格品相互独立。

(1)记 20 件产品中恰有 2 件不合格品的概率为 $f(p)$，求 $f(p)$的最大值点 p_0。

(2)现对一箱产品检验了 20 件，结果恰有 2 件不合格品，以(1)中确定的 p_0 作为 p 的值，已知每件产品的检验费用为 2 元，若有不合格品进入用户手中，则工厂要对每件不合格品支付 25 元的赔偿费用。

①若不对该箱余下的产品检验，这一箱产品的检验费用与赔偿费用的和记为 X，求 $E(X)$；

②以检验费用与赔偿费用和的期望值为决策依据，是否该对这箱余下的所有产品作检验？

解析：(1)记 20 件产品中恰有 2 件不合格品的概率为 $f(p)$，则 $f(p)=C_{20}^2p^2(1-p)^{18}$。

故 $f'(p)=C_{20}^2[2p(1-p)^{18}-18p^2(1-p)^{17}]=2C_{20}^2p(1-p)^{17}(1-10p)$。

令 $f'(p)=0$，得 $p=0.1$，

当 $p\in(0,0.1)$时，$f'(p)>0$；

当 $p\in(0.1,1)$时，$f'(p)<0$。

故 $f(p)$的最大值点 $p_0=0.1$。

(2)①由(1)知 $p=0.1$。

令 Y 表示余下的 180 件产品中的不合格品数，依题意知 $Y\sim B(180,0.1)$。

$X=20\times2+25Y$，即 $X=40+25Y$。

$E(X)=E(40+25Y)=40+25E(Y)=40+25\times180\times0.1=490$。

②如果对余下的产品检验，这一箱产品所需要的检验费为 400 元。

因为 $E(X)=490>400$，所以应该对余下的产品进行检验。

点评：题目当中出现了“各件产品是否为不合格品相互独立”的字眼，符合二项分布的特征。

(责任编辑 徐利杰)