超几何分布

zcl.space

目录

1 定义

设一个坛子里一共有N个球，其中m个白球，$N-m$个黑球，从中随机的（无放回）取出n个球，令X表示取出来的白球数，那么：

$$P\{X=i\}=\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}},\quad i=0,1,\ldots,n \tag{1.1}$$

一个随机变量X，如果其概率分布满足式 (1.1)就说X服从超几何分布。

2 超几何分布的期望和方差

我们采用和 二项分布 类似的计算方法。

$$\begin{aligned}
E[X^k] &= \sum_{i=0}^{n} i^k P\{X=i\} && (2.1)\\
&= \sum_{i=1}^{n} i^k \binom{m}{i}\binom{N-m}{n-i}\Big/\binom{N}{n} && (2.2)\\
&= \frac{mn}{N}\sum_{i=1}^{n} i^{k-1}\binom{m-1}{i-1}\binom{N-m}{n-i}\Big/\binom{N-1}{n-1} && (2.3)\\
&= \frac{mn}{N}\sum_{j=0}^{n-1}(j+1)^{k-1}\frac{\binom{m-1}{j}\binom{N-m}{n-1-j}}{\binom{N-1}{n-1}} && (2.4)\\
&= \frac{mn}{N}E[(Y+1)^{k-1}] && (2.5)
\end{aligned}$$

其中Y是参数为$(n-1, N-1, m-1)$的超几何随机变量。在上式的求解过程中，我们用到了两个恒等式：

$$\begin{aligned}
i\binom{m}{i} &= m\binom{m-1}{i-1} && (2.6)\\
n\binom{N}{n} &= N\binom{N-1}{n-1} && (2.7)
\end{aligned}$$

对于式 (2.1)，我们令$k=1$，所以有：

$$E[X] = \binom{mn}{N} \tag{2.8}$$

换句话说，如果从N个球（其中有m个白球）中随机抽取n个，那么其中白球数的期望为mn/N。

对于式 (2.1)，令$k=2$，则有：

$$\begin{aligned}
E[X^2] &= \frac{mn}{N}E[Y+1] && (2.9)\\
&= \frac{mn}{N}\left[\frac{(n-1)(m-1)}{N-1}+1\right] && (2.10)
\end{aligned}$$

另外$E[X]=mn/N$，所以我们有：

$$\mathrm{Var}(X) = \frac{mn}{N}\left[\frac{(n-1)(m-1)}{N-1}+1-\frac{mn}{N}\right] \tag{2.11}$$

令$p=m/N$，并利用恒等式：

$$\frac{m-1}{N-1} = \frac{Np-1}{N-1} = p-\frac{1-p}{N-1} \tag{2.12}$$

得到：

$$\mathrm{Var}(X) = np(1-p)(1-\tfrac{n-1}{N-1}) \tag{2.13}$$

对于超几何分布，我们知道期望为$\frac{mn}{N} = np, p = \frac{m}{N}$,$p$是白球的比例。当$N$很大时，观察 (2.13)，我们有：

$$\mathrm{Var}[X] \approx np(1-p) \tag{2.14}$$

回忆二项分布，我们很容易发现其和超几何分布的相似性。

3 例子

栖息于某个地区的动物个体总数为N，为了得到这个N的大致估计，神态学家常常做这样的试验：先捉住m个，然后打上标签，放回大自然。过一段时间，等这m个动物充分分散到其他动物中的时，再捕捉n个。假设X为第二批捕捉的n个动物中带标记的个数。如果前后两次捕捉过程中动物的总数没有发生变化，而且捉住每一只动物的可能性是一样的，那么X为一超几何随机变量，满足：

$$P\{X=i\} = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} = P_i(N) \tag{3.1}$$

现在假设i为X的观测值，那么因为$P_i(N)$表示该地区事实上总共有N个动物的条件下观测时间X的取值的概率，故使$P_i(N)$达到最大值的N值应当是动物个体总数N的一个合理估计。这样的估计称为极大似然估计估计。

求$P_i(N)$最大值的最简单的方法是：首先注意

$$\frac{P_i(N)}{P_i(N-1)} = \frac{(N-m)(N-n)}{N(N-m-n+i)} \tag{3.2}$$

使上式中的比值大于1，则有：

$$(N-m)(N-n) \geq N(N-m-n+i) \tag{3.3}$$

或者必须有：

$$N \leq \frac{mn}{i} \tag{3.4}$$

所以$P_i(N)$值是先上升然后下降。且在不超过mn/i的最大整数处达到其最大值。这个最大整数就是N的最大似然估计。

上述估计还可以这样求得：假设在这个地区内有标记的动物所占的比例为m/N，应当近似的等于第二次捕捉的动物中做过标记的动物所占的比例i/n。

4 超几何分布和二项分布的关系

从N个球（白球比例为$p = m/N$）中，无放回随机抽取n个球，那么取中的白球数为超几何分布随机变量。如果对于n来讲，m和N都很大，那么有放回和无放回取球没什么差别，因为当m和N很大时，不管前面取了哪个球，接下来的取到的失败求的概率任然近似于p。换言之，当m和N相比n很大时，X的分布列应该近似等于参数为(n, p)的二项随机变量的分布列。为了证明这个直觉，注意，如果X是超几何分布，那么对于$i \le n$有：

$$\begin{aligned}
P\{X = i\} &= \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} && (4.1)\\
&= \frac{m!}{(m-i)!i!}\frac{(N-m)!}{(N-m-n+i)!(n-i)!}\frac{(N-n)!n!}{N!} && (4.2)\\
&= \binom{n}{i}\frac{m}{N}\frac{m-1}{N-1}\cdots\frac{m-i+1}{N-i+1}\frac{N-m}{N-i}\frac{N-m-1}{N-i-1}\cdots\frac{N-m-(n-i-1)}{N-i-(n-i-1)} && (4.3)\\
&= \binom{n}{i}p^i(1-p)^{n-i} && (4.4)
\end{aligned}$$

其中最后一个等式成立的条件时$p = m/N$且m, N相对于n和i都很大。我们在求超几何变量的期望和方差时也验证了这个结论。