详细解答

STEP II 2000，第 13 题

概率与统计 · 捕捉再捕捉估计

STEP II 2000，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/00-s2-q13

本题在 *Beyond STEP* 第 13.2 节计数田鼠中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch13/voles.html

关键思路

这是捉放法（capture–recapture）：第二次抽样是从 N 只田鼠中随机选出 200 只，而当中有 200 只带有标记，所以样本中带标记田鼠的数目服从超几何分布（hypergeometric distribution）。要找出 p_N 在何处最大，不要求导（N 的阶乘（factorial）并没有简单意义下的导数）；应该改为通过比值 p_N/p_{N-1} 比较相邻两项，这时几乎所有东西都会约去。最后的恒等式则不费吹灰之力：第三次抽样的各个概率之和必定是 1。

解答

概率 p_N

把第二次抽样看成从 N 只田鼠中选出 200 只，每个大小为 200 的子集出现的机会均等。这样的子集共有 $\binom{N}{200}$ 个。在 N 只田鼠中，200 只带有标记，$N-200$ 只没有。一个恰好含 11 只带标记田鼠的样本，就是选出 11 只带标记的和 189 只没有标记的，所以

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

把含有 N 的两个二项式系数（binomial coefficient）写成阶乘，

$$p_N = \binom{200}{11} \cdot \frac{(N-200)!}{189!\,(N-389)!} \cdot \frac{200!\,(N-200)!}{N!}$$
$$= k\,\frac{((N-200)!)^2}{N!\,(N-389)!}, \qquad \text{其中} k = \binom{200}{11}\frac{200!}{189!},$$

而这个常数与 N 无关。这正是所要求的形式。（这里需要 $N \geq 389$，因为两次抽样合共包含 $200 + 189 = 389$ 只不同的田鼠。）

求 p_N 的最大值

对于 $N \geq 390$，把 p_N 除以 p_{N-1}。利用 $(N-200)! = (N-200)\,(N-201)!$、$N! = N\,(N-1)!$ 和 $(N-389)! = (N-389)\,(N-390)!$，所有东西都会约去，只余下

$$\frac{p_N}{p_{N-1}} = \frac{(N-200)^2}{N(N-389)}.$$

分子和分母都是正数，所以

$$\begin{aligned} p_N > p_{N-1} &\iff (N-200)^2 > N(N-389) \\ &\iff N^2 - 400N + 40000 > N^2 - 389N \\ &\iff 40000 > 11N \\ &\iff N < 3636.\overline{36}. \end{aligned}$$

想一想 这告诉我们数列（sequence）$p_{389}, p_{390}, \ldots$ 的形状是怎样的?

当 $390 \leq N \leq 3636$ 时，比值大于 1，所以数列严格递增，直至 p_{3636}。当 $N \geq 3637$ 时，比值小于 1（等号成立需要 $11N = 40000$，而这没有整数解），所以数列自此严格递减。因此 p_N 有唯一的最大值，位于

$$N = 3636.$$

这就是 $\frac{200\times200}{11}$ 的整数部分。这也正是你会猜到的答案：只要认为样本中带标记的比例 $\frac{11}{200}$ 应该等于总体中带标记的比例 $\frac{200}{N}$ 便可。

第三次抽样

取总体大小为 3636，即刚才求得的估计值。第二次放回之后，带标记的田鼠是原来的 200 只加上新标记的 189 只，所以有 389 只带有标记，$3636 - 389 = 3247$ 只没有标记。按照与之前完全相同的推理，一个大小为 200 的随机样本中恰好有 j 只带标记田鼠的概率是

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}, \qquad j = 0, 1, \ldots, 200.$$

由于 $389 \geq 200$ 且 $3247 \geq 200$，由 0 至 200 的每一个 j 都有可能出现，而这些就是全部可能的结果。它们互斥（mutually exclusive），所以它们的概率之和是 1：

$$\sum_{j=0}^{200} \frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}} = 1.$$

乘以 $\binom{3636}{200}$ 便得

$$\sum_{j=0}^{200} \binom{389}{j}\binom{3247}{200-j} = \binom{3636}{200},$$

这正是所要证明的。

验算

- 用整数二项式系数准确计算 p_N：对于若干个 N 的值（例如 389、1000 和 5000），结果与阶乘公式一致；而在 $389 \leq N < 20000$ 的范围内搜索，最大值位于 $N = 3636$。
- 在转折点处，$p_{3636}/p_{3635} = 3436^2/(3636 \times 3246) \approx 1.0003 > 1$，但 $p_{3637}/p_{3636} < 1$。
- 最后的恒等式两边以整数准确计算，结果相等。留意 $389 + 3247 = 3636$，这是必然的。

常见错误

- 把 N 当作连续变量来求导。应该改用相邻两项的比值，并准确地判定不等式。
- 忘记了第二次放回之后，带标记的田鼠是 389 只，而不是 200 只。
- 说“最大值位于 $N = 3636.36$”：总体大小必须是整数，而且你应该检查相邻的两个整数中哪一个胜出。

STEP 之外

选取使观测数据出现的概率最大的参数值，就是*最大似然法*（maximum likelihood），这是大学一年级统计学的核心思想。在这里，它给出捉放法的 Lincoln–Petersen 估计量（estimator）。最后的恒等式是范德蒙德恒等式（Vandermonde's identity），而所用的论证方法，即用两种方法数算同一样东西，是组合数学（combinatorics）中一种强而有力的技巧。Grimmett and Welsh 的 *Probability: An Introduction*（概率入门教材）涵盖了超几何分布和这类论证方法。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。