13.2 點算田鼠

工作紙

題目

STEP II 2000，第 13 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：

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

提示

請先自行嘗試解題。如果遇到困難，請逐一閱讀提示，每讀完一個便回到題目再試。標有停一停，試一試的提示是關鍵步驟。每個提示之後均留有空位供你演算。

> **想一想**　種羣共有 N 隻田鼠，其中 200 隻帶有記號。現隨機抽取 200 隻作為樣本。恰好有 11 隻帶有記號的概率是多少？

階段 1　似然（likelihood）

涵蓋：概率 p_N

> **想一想**　在所有等可能的 200 隻樣本（sample）之中，從有標記的田鼠中選出 11 隻，再從沒有標記的田鼠中選出 189 隻。然後把含有 N 的二項式係數（binomial coefficient）寫成階乘（factorial）的形式。

階段 2 對整數求最大值

涵蓋：求 p_N 的最大值

> **停一停，試一試** N 是整數，所以應比較 p_{N+1} 與 p_N（或 p_N 與 p_{N-1}）。大部分階乘都會約去。這個比值何時大於 1？

階段 3　第三個樣本

涵蓋：第三個樣本

? **想一想**　抽取第二個樣本之後，有多少隻田鼠帶有標記？按照你對 N 的估計，有多少隻沒有標記？寫出第三個 200 隻的樣本中恰好有 j 隻有標記田鼠的概率。這些概率加起來等於甚麼？

背後的數學

最大似然

最大似然原理由費雪（Ronald Fisher）在 1910 至 1920 年代發展出來，內容是：以使觀察到的數據出現概率最大的那個值，作為未知參數的估計。它是至今應用最廣的估計方法，由以曲線擬合數據，到訓練機器學習中的神經網絡，無一不用；神經網絡的「損失函數」（loss function）通常就是負對數似然。費雪證明了，對於大樣本，最大似然估計值近似服從正態分佈（normal distribution），而且其精確程度不遜於任何合理的估計量（estimator），前提是模型具備某些良好性質，而這些性質對大多數日常模型都成立。不過也有重要的例外。在下一節（Section 13.3）中，未知量是可能出現的最大值，即分佈的一個端點，在那裏最大似然估計值的表現大不相同：它的誤差並不近似服從正態分佈，而且誤差是以 $\frac{1}{n}$ 而非 $\frac{1}{\sqrt{n}}$ 的速度縮小。

這個估計有多精確？

圖 13.2.1（見網上該節的完整版）顯示似然曲線很寬闊：11 次重捕所提供的信息並不多。粗略計算一下：第二個樣本中有記號田鼠的數目近似服從泊松分佈（Poisson distribution），平均值為 $\frac{200\times 200}{N}$，所以其標準差（standard deviation）約為 $\sqrt{11} \approx 3.3$，相對誤差約為 30%，而這個誤差會傳遞到 N 的估計上。若要把 N 的估計誤差控制在 10% 以內，生物學家需要大約 100 次重捕。統計學家還會修正這個簡單估計的輕微偏差（bias），採用查普曼估計量（Chapman estimator）$\frac{(M+1)(C+1)}{R+1} - 1$，其中 M 和 C 是兩次捕捉的

數量，R 是重捕到的數目。

拉普拉斯點算法國人口

拉普拉斯（Laplace）在 1802 年用了同樣的想法，在沒有人口普查的情況下估計法國的人口。他知道全國一年內登記的出生人數，並且就一批抽樣的市鎮而言，同時知道其人口和出生人數。他假設人口與出生人數之比在各地都相同，從而估計法國人口約為 28 百萬，甚至給出了不確定程度的度量。後來這個方法在動物種羣研究中被重新發現：C. G. J. Petersen 在 1890 年代於丹麥的峽灣為鰈魚做記號，Frederick Lincoln 則在 1930 年代使用鳥環；這個方法至今仍常被稱為林肯—彼得森（Lincoln–Petersen）估計量。

延伸練習

1. **一般公式。**若有 M 隻動物做了記號，而第二個含 C 隻動物的樣本中有 R 隻帶有記號，證明 N 的最大似然估計值是 $\frac{MC}{R}$ 的整數部分。

2. **范德蒙恆等式。**考慮從 a 名女士和 b 名男士中選出 c 人組成委員會的方法數目，以此證明 $\sum_j \binom{a}{j}\binom{b}{c-j} = \binom{a+b}{c}$。

3. **擲硬幣。**一枚硬幣擲出正面的概率 p 未知，擲 n 次後觀察到 h 次正面。證明 p 的最大似然估計值是 $\frac{h}{n}$。

4. **沒有重捕。**如果第二個樣本中沒有任何帶記號的田鼠，最大似然估計值會變成怎樣？這說明了這個方法的甚麼特點？

延伸閱讀

- **David Spiegelhalter, *The Art of Statistics***（Pelican），談估計與不確定性。
- **George Seber, *The Estimation of Animal Abundance***（Blackburn）。標記重捕法的標準參考書。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。數理統計的入門課本，內容包括最大似然。

完整的引導式解答見第 13.2 節：

https://beyondstep.keithmcnulty.org/zh-hant/ch13/voles.html

每星期一道 STEP 題目，完整講解：The STEP Up，以電郵免費發送，訂閱網址：keithmcnulty.substack.com