詳細解答

STEP I 1995，第 14 題

概率與統計 • 比例的估計

> **STEP I 1995，第 14 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/95-s1-q14

關鍵思路

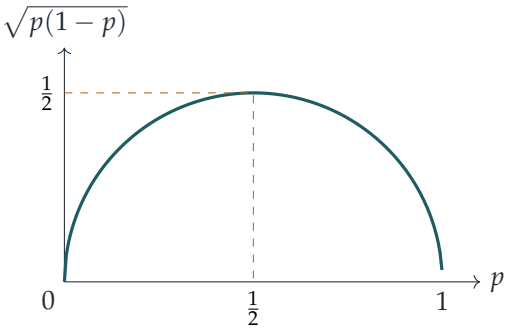
樣本比例（sample proportion）就是一個二項（binomial）計數除以 n。當 n 很大時，它近似服從正態分佈（normal distribution），平均值為 p，方差（variance）為 $p(1-p)/n$。麻煩在於，這個方差偏偏取決於我們不知道的那個 p。第 (i) 部分正好解決這個問題：它給出標準差（standard deviation）最壞情況下的值，所以按最壞情況選定的樣本大小，無論 p 是多少都適用。第 (iii) 部分接着問：如果我們對 p 已經有粗略的認識，情況會有甚麼不同。

解答

第 (i) 部分

用配方法（completing the square），設 $f(p) = p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2$。所以在 $[0,1]$ 上 $f(p) \le \frac{1}{4}$，等號僅在 $p = \frac{1}{2}$ 時成立。平方根是遞增函數，因此

$$\max_{0 \le p \le 1} \sqrt{p(1-p)} = \sqrt{\tfrac{1}{4}} = \tfrac{1}{2}, \qquad \text{取於} p = \tfrac{1}{2}.$$



第 (ii) 部分

設 X 為樣本中女性的人數。每個被抽中的人是女性的概率為 p。我們假設總體（population）很大，因此這 n 次抽選實際上可視為互相獨立（從龐大的總體中不放回抽樣，與放回抽樣幾乎沒有分別）。於是

$$X \sim \mathrm{B}(n,p), \qquad \hat{p} = \frac{X}{n}, \qquad \mathrm{E}(\hat{p}) = p, \qquad \mathrm{Var}(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

當 n 很大時，由中心極限定理（central limit theorem，即二項分佈的正態近似）可得

$$\hat{p} \approx \mathrm{N}\left(p, \frac{p(1-p)}{n}\right), \qquad \text{所以} \qquad Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \approx \mathrm{N}(0,1).$$

我們要求 $\mathrm{P}(|\hat{p}-p| > 0.01) < 0.01$，即

$$\mathrm{P}\left(|Z| > \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) < 0.01.$$

由對稱性，這 1% 平均分到兩邊尾部，每邊 0.5%。由於 $\Phi(2.576) = 0.995$，我們需要

$$\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}} > 2.576 \quad \Longleftrightarrow \quad \sqrt{n} > 257.6\sqrt{p(1-p)}.$$

想一想 我們不知道 p，那麼怎樣從一個含有它的不等式去選定 n 呢？

這正是第 (i) 部分派上用場的地方。無論 p 是多少，都有 $\sqrt{p(1-p)} \le \frac{1}{2}$。所以只要 $\sqrt{n} > 257.6 \times \frac{1}{2} = 128.8$，不等式便對任何 p 都成立。這個條件也是必要的，因為當 $p = \frac{1}{2}$ 時（而女性所佔的比例確實接近這個值），上界可以取到。因此我們需要

$$n > 128.8^2 = 16\,589.44,$$

所求的樣本大小為

$$\boxed{n = 16\,590} \qquad (\text{約}16\,600).$$

若採用更精確的百分點 $z = 2.5758$，則得 $n > 16\,587.2$。無論用哪個值，答案都是「約 16 600」，而第三位有效數字則視乎所用的統計表而定。

第 (iii) 部分

一般而言，所需的條件是

$$n > \left(\frac{2.576}{0.01}\right)^2 p(1-p) \approx 66\,360\, p(1-p).$$

16 590 這個數值來自最壞情況 $p = \frac{1}{2}$。如果我們知道 p 離 $\frac{1}{2}$ 相當遠，便可以為 $p(1-p)$ 採用一個較小的上界。

(a) 左撇子。大約 10% 的人是左撇子。如果可以穩妥地假設 $p \le 0.15$，那麼 $p(1-p) \le 0.15 \times 0.85 = 0.1275$，因為 $p(1-p)$ 在 $[0, \frac{1}{2}]$ 上遞增。由此得 $n > 8461$。如果恰好 $p = 0.1$，則 $n > 66\,360 \times 0.09 \approx 5970$，即大約 6000。無論哪種情況，所需樣本都遠小於一無所知時所需的 16 590。正態近似仍然相當準確，因為 np 數以百計。

(b) 百萬富翁。這時 p 非常小，也許介乎 0.001 至 0.01 之間。按字面計算，$p \le 0.01$ 給出 $p(1-p) \le 0.0099$ 及 $n > 657$。但這樣做並沒有抓住重點：

- 當 p 本身只有大約 0.001 時，0.01 的誤差毫無用處：估計值 $\hat{p} = 0$ 與真實值的差距已經在 0.01 之內。我們真正想要的是相對準確度，例如誤差不超過 1% 的 p。這樣便需要 $n > 2.576^2(1-p)/(p \times 0.01^2)$，當 $p = 0.001$ 時約為 6.6×10^7。這比先前的數字大得多。
- 當 p 極小而 n 只有數百時，np 小於 1。這時二項分佈非常偏斜，正態近似不再有效；以泊松近似（Poisson approximation）作為模型會更合適。
- 實際上，百萬富翁很難接觸得到，而且人們未必會如實說出自己的財富，所以能否做到隨機抽樣、受訪者是否誠實作答，兩者都成疑問。

因此，就左撇子而言，粗略知道 p 的大小可以讓我們採用較小的樣本。就百萬富翁而言，任何合理的準確度要求都需要大得多的樣本，或者要改用完全不同的抽樣方法。

驗算

- 當 $p = \frac{1}{2}$ 而 $n = 16\,590$ 時，正態近似給出 $\mathrm{P}(|\hat{p} - p| > 0.01) = 2\big(1 - \Phi(2.5762)\big) \approx 0.00999$，剛好略低於 1%。
- 把 $n = 16\,590$、$p = \frac{1}{2}$ 時的精確二項概率加起來，結果約為 0.0102。稍微超出的部分源於二項分佈的離散性（這裏的正態近似沒有作連續性修正（continuity correction））。當 $p = 0.3$ 時，精確值約為 0.005，綽綽有餘地符合要求，正如第 (i) 部分所預示。
- 當 $p = 0.1$ 而 $n = 6000$ 時，精確的二項概率約為 0.0098，印證了 (iii)(a) 中的數字。

常見錯誤

- 用了 1.96 或 2.326 而不是 2.576。這 1% 是分到兩邊尾部的，所以我們需要 $\Phi(z) = 0.995$。
- 把 n 留作以 p 表示的式子。題目要的是一個對任何 p 都適用的確定數值，這正是設置第 (i) 部分的原因。
- 把「$\sqrt{p(1-p)} \leq \frac{1}{2}$」理解成得出 $n > 128.8$。必須超過 128.8 的是 $\sqrt{n}$，而不是 n。

STEP 之外

民意調查中所報的「誤差範圍」(margin of error)，背後就是這個標準計算。一項訪問約 1000 人的民調，在 95% 置信水平下的誤差大約為 ± 3 個百分點，用的正是第 (i) 部分的最壞情況論證。到了大學，這會引伸出置信區間（confidence interval），以及如何估計罕見比例的問題；在那種情況下，泊松方法和精確二項方法會取代正態近似。Blitzstein 與 Hwang 合著的 *Introduction to Probability*（概率論入門）對中心極限定理和這些近似方法有十分清晰易讀的講解。

Beyond STEP 詳細解答：https://beyondstep.keithmcnulty.org/zh-hant/。解答文字採用 CC0 1.0；題目本身版權屬 Cambridge University Press & Assessment 所有（©），本文件不予轉載。由機器從英文翻譯而成；尚未經審閱。