详细解答

STEP I 1995，第 14 题

概率与统计 • 比例的估计

STEP I 1995，第 14 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/95-s1-q14

关键思路

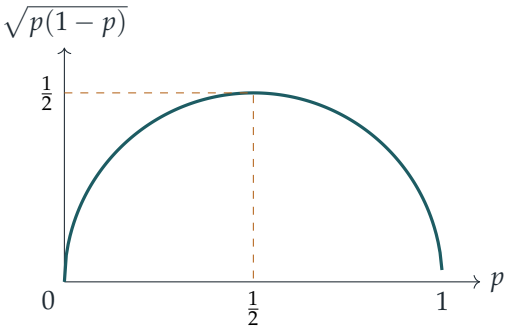
样本比例（sample proportion）就是一个二项（binomial）计数除以 n。当 n 很大时，它近似服从正态分布（normal distribution），平均值为 p，方差（variance）为 $p(1-p)/n$。麻烦在于，这个方差偏偏取决于我们不知道的那个 p。第 (i) 部分正好解决这个问题：它给出标准差（standard deviation）最坏情况下的值，所以按最坏情况选定的样本大小，无论 p 是多少都适用。第 (iii) 部分接着问：如果我们对 p 已经有粗略的认识，情况会有甚么不同。

解答

第 (i) 部分

用配方法（completing the square），设 $f(p) = p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2$。所以在 $[0,1]$ 上 $f(p) \le \frac{1}{4}$，等号仅在 $p = \frac{1}{2}$ 时成立。平方根是递增函数，因此

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



第 (ii) 部分

设 X 为样本中女性的人数。每个被抽中的人是女性的概率为 p。我们假设总体(population)很大，因此这 n 次抽选实际上可视为互相独立（从庞大的总体中不放回抽样，与放回抽样几乎没有分别）。于是

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

当 n 很大时，由中心极限定理（central limit theorem，即二项分布的正态近似）可得

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

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

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

由对称性，这 1% 平均分到两边尾部，每边 0.5%。由于 $\Phi(2.576) = 0.995$，我们需要

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

想一想 我们不知道 p，那么怎样从一个含有它的不等式去选定 n 呢?

这正是第 (i) 部分派上用场的地方。无论 p 是多少，都有 $\sqrt{p(1-p)} \le \frac{1}{2}$。所以只要 $\sqrt{n} > 257.6 \times \frac{1}{2} = 128.8$，不等式便对任何 p 都成立。这个条件也是必要的，因为当 $p = \frac{1}{2}$ 时（而女性所占的比例确实接近这个值），上界可以取到。因此我们需要

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

所求的样本大小为

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

若采用更精确的百分点 $z = 2.5758$，则得 $n > 16\,587.2$。无论用哪个值，答案都是“约 16 600”，而第三位有效数字则视乎所用的统计表而定。

第 (iii) 部分

一般而言，所需的条件是

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

16 590 这个数值来自最坏情况 $p = \frac{1}{2}$。如果我们知道 p 离 $\frac{1}{2}$ 相当远，便可以为 $p(1-p)$ 采用一个较小的上界。

(a) 左撇子。大约 10% 的人是左撇子。如果可以稳妥地假设 $p \le 0.15$，那么 $p(1-p) \le 0.15 \times 0.85 = 0.1275$, 因为 $p(1-p)$ 在 $[0, \frac{1}{2}]$ 上递增。由此得 $n > 8461$。如果恰好 $p = 0.1$, 则 $n > 66\,360 \times 0.09 \approx 5970$，即大约 6000。无论哪种情况，所需样本都远小于一无所知时所需的 16 590。正态近似仍然相当准确，因为 np 数以百计。

(b) 百万富翁。这时 p 非常小，也许介乎 0.001 至 0.01 之间。按字面计算，$p \le 0.01$ 给出 $p(1-p) \le 0.0099$ 及 $n > 657$。但这样做并没有抓住重点：

- 当 p 本身只有大约 0.001 时，0.01 的误差毫无用处：估计值 $\hat{p} = 0$ 与真实值的差距已经在 0.01 之内。我们真正想要的是相对准确度，例如误差不超过 1% 的 p。这样便需要 $n > 2.576^2(1-p)/(p \times 0.01^2)$，当 $p = 0.001$ 时约为 6.6×10^7。这比先前的数字大得多。
- 当 p 极小而 n 只有数百时，np 小于 1。这时二项分布非常偏斜，正态近似不再有效；以泊松近似（Poisson approximation）作为模型会更合适。
- 实际上，百万富翁很难接触得到，而且人们未必会如实说出自己的财富，所以能否做到随机抽样、受访者是否诚实作答，两者都成疑问。

因此，就左撇子而言，粗略知道 p 的大小可以让我们采用较小的样本。就百万富翁而言，任何合理的准确度要求都需要大得多的样本，或者要改用完全不同的抽样方法。

验算

- 当 $p = \frac{1}{2}$ 而 $n = 16\,590$ 时，正态近似给出 $\mathrm{P}(|\hat{p} - p| > 0.01) = 2\big(1 - \Phi(2.5762)\big) \approx 0.00999$，刚好略低于 1%。
- 把 $n = 16\,590$、$p = \frac{1}{2}$ 时的精确二项概率加起来，结果约为 0.0102。稍微超出的部分源于二项分布的离散性（这里的正态近似没有作连续性修正（continuity correction））。当 $p = 0.3$ 时，精确值约为 0.005，绰绰有余地符合要求，正如第 (i) 部分所预示。
- 当 $p = 0.1$ 而 $n = 6000$ 时，精确的二项概率约为 0.0098，印证了 (iii)(a) 中的数字。

常见错误

- 用了 1.96 或 2.326 而不是 2.576。这 1% 是分到两边尾部的，所以我们需要 $\Phi(z) = 0.995$。
- 把 n 留作以 p 表示的式子。题目要的是一个对任何 p 都适用的确定数值，这正是设置第 (i) 部分的原因。
- 把“$\sqrt{p(1-p)} \leq \frac{1}{2}$”理解成得出 $n > 128.8$。必须超过 128.8 的是 $\sqrt{n}$，而不是 n。

STEP 之外

民意调查中所报的“误差范围”（margin of error），背后就是这个标准计算。一项访问约 1000 人的民调，在 95% 置信水平下的误差大约为 ± 3 个百分点，用的正是第 (i) 部分的最坏情况论证。到了大学，这会引伸出置信区间（confidence interval），以及如何估计罕见比例的问题；在那种情况下，泊松方法和精确二项方法会取代正态近似。Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）对中心极限定理和这些近似方法有十分清晰易读的讲解。

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。