详细解答

STEP II 2024，第 11 题

概率与统计 · 分组检测的期望值

STEP II 2024，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2024 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

www.ocr.org.uk/Images/725753-2024-past-papers.zip

关键思路

只有当混合样本的检测通常呈阴性时，合并检测（pooling）才划算；而每人的期望（expected）检测次数原来是 $1+\frac{1}{k}-(1-p)^k$。要问这个数何时不超过 1，就变成比较 $(1-p)$ 与 $k^{-1/k}$ 的大小，这正是第 (i) 部分要讨论 $x^{1/x}$ 的原因：当 p 较大时，最佳的组别人数就是使 $k^{1/k}$ 最大的那个整数。

解答

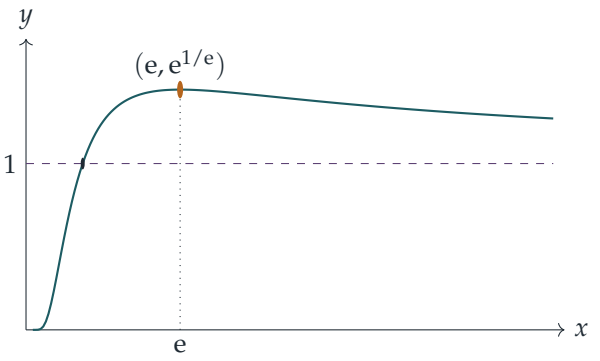
第 (i) 部分

对 $x>0$，记 $y=x^{1/x}=\mathrm{e}^{(\ln x)/x}$，于是 $\ln y=\dfrac{\ln x}{x}$。求导（differentiating）得

$$\frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{1-\ln x}{x^2}.$$

由于 $y>0$，$\frac{\mathrm{d}y}{\mathrm{d}x}$ 与 $1-\ln x$ 同号：当 $0<x<\mathrm{e}$ 时为正，在 $x=\mathrm{e}$ 时为零，当 $x>\mathrm{e}$ 时为负。因此曲线只有一个转折点（turning point），即位于 $(\mathrm{e},\mathrm{e}^{1/\mathrm{e}})\approx(2.72,\,1.44)$ 的极大值（maximum）点。

至于图像两端：当 $x\to0^+$ 时，$(\ln x)/x\to-\infty$，所以 $y\to0$；当 $x\to\infty$ 时，$(\ln x)/x\to0$，所以 $y\to1$，且从上方趋近（因为对 $x>1$ 有 $\ln y>0$）。曲线经过点 $(1,1)$。



对正整数 n 而言，此函数在 $(0, \mathrm{e}]$ 上递增，在 $[\mathrm{e}, \infty)$ 上递减，所以 $1^{1/1} < 2^{1/2}$ 且 $3^{1/3} > 4^{1/4} > 5^{1/5} > \cdots$。因此最大值是 $2^{1/2}$ 或 $3^{1/3}$。把两者都升到六次方，得 $2^3 = 8$ 和 $3^2 = 9$，所以 $3^{1/3} > 2^{1/2}$，且

$$\max_n n^{1/n} = 3^{1/3}.$$

第 (ii) 部分

记 $q = 1 - p$。考虑一组 k 人。由独立性（independence），组内无人带有该酶的概率是 q^k，而只有在这种情况下混合样本才呈阴性。所以该组所用的检测次数为

$$\begin{cases} 1 \text{ 次检测} & \text{概率为} q^k, \\ 1 + k \text{ 次检测} & \text{概率为} 1 - q^k, \end{cases}$$

其期望检测次数是 $1 + k(1 - q^k)$。共有 $r = N/k$ 组，而期望值具可加性，所以期望总次数为

$$E = \frac{N}{k}(1 + k(1 - q^k)) = N\left(1 + \frac{1}{k} - (1 - p)^k\right).$$

第 (iii) 部分

由第 (ii) 部分，$E \le N$ 当且仅当

$$\frac{1}{k} \le (1 - p)^k \iff 1 - p \ge k^{-1/k} = \frac{1}{k^{1/k}},$$

这里取了正的 k 次方根（两边均为正数，而 $t \mapsto t^{1/k}$ 是递增的）。

> ? **想一想** 哪个 k 容许最大的 p?

对给定的 k，条件是 $p \le 1 - k^{-1/k}$，而当 $k^{1/k}$ 最大时，这个上界也最大。由第 (i) 部分，对每个整数 $k > 1$ 都有 $k^{1/k} \le 3^{1/3}$，所以无论选哪个 k，都必须有 $p \le 1 - 3^{-1/3}$。反过来，

N 是 3 的倍数，所以 $k = 3$ 是容许的组别人数，而取 $k = 3$ 时条件正好是 $p \le 1 - 3^{-1/3}$（等号成立时 $E = N$）。因此最大值为

$$p = 1 - 3^{-1/3} \approx 0.307.$$

要证明这个值大于 $\frac{1}{4}$：

$$1 - 3^{-1/3} > \tfrac{1}{4} \iff 3^{-1/3} < \tfrac{3}{4} \iff \tfrac{1}{3} < \tfrac{27}{64} \iff 64 < 81,$$

其中中间一步两边取立方（对正数而言，取立方保持大小次序）。最后一个式子成立，所以 $1 - 3^{-1/3} > \frac{1}{4}$。

第 (iv) 部分

由二项式展开（binomial expansion），

$$(1-p)^k = 1 - kp + \binom{k}{2}p^2 - \cdots.$$

若 pk 很小，则 $\binom{k}{2}p^2 < \frac{1}{2}(pk)^2$ 更小得多，之后各项亦然，所以 $(1-p)^k \approx 1 - pk$。代入第 (ii) 部分的结果，

$$E \approx N\left(1 + \frac{1}{k} - (1 - pk)\right) = N\left(\frac{1}{k} + pk\right).$$

（事实上，伯努利不等式（Bernoulli's inequality）$(1-p)^k \ge 1 - pk$ 表明这个近似永远不会低估 E；而把展开式多保留一项，则可证明误差最多为 $\frac{1}{2}(pk)^2 N$。）

取 $p = 0.01$ 和 $k = 10$，得 $pk = 0.1$，确实很小，而

$$E \approx N\left(\frac{1}{10} + 0.1\right) = 0.2N,$$

所以这个方案只需约 20% 的检测次数，相比之下逐一检测每个人需要 N 次。

验算

- 取 $p = 0.01$ 和 $k = 10$ 作精确计算：$E/N = 1.1 - 0.99^{10} \approx 1.1 - 0.9044 = 0.1956$，约为 19.6%，与近似值之差在误差上界 $\frac{1}{2}(0.1)^2 = 0.005$ 之内。
- 当 $p = 1 - 3^{-1/3}$ 且 $k = 3$ 时：$(1-p)^3 = \frac{1}{3}$，所以 $E = N(1 + \frac{1}{3} - \frac{1}{3}) = N$，刚好落在边界上。
- 检查第 (ii) 部分是否合理：当 $p \to 0$ 时，$E \to N/k$（每组一次检测）；当 $p \to 1$ 时，$E \to N(1 + \frac{1}{k})$（每组都要重新逐一检测）。此公式亦已用电脑逐一列举小组别的所有结果加以验证。

常见错误

- 第 (i) 部分答 $e^{1/e}$：n 必须是整数，所以应比较 $2^{1/2}$ 与 $3^{1/3}$。
- 第 (iii) 部分只验证了 $k = 3$，却没有解释为何其他组别人数不可能容许更大的 p。
- 在呈阳性的组别中漏计混合检测本身，得出 k 次而非 $k + 1$ 次检测。

STEP 之外

这就是 Dorfman 的群组检测（group testing）方案，它在第二次世界大战期间为筛查士兵的血液样本而设计，后来在 COVID-19 检测中再度被采用。把 $\frac{1}{k} + pk$ 最小化，得 $k \approx 1/\sqrt{p}$（所以这里 $k = 10$ 是最佳选择），而所需检测次数约为 $2\sqrt{p}\,N$；信息理论（information theory）中研究的更巧妙的自适应（adaptive）方案，效果还要更好。关于背后的概率理论，可参阅 Blitzstein 与 Hwang 的 *Introduction to Probability*（《概率导论》）。

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。