详细解答

STEP II 1996，第 14 题

概率与统计 • 矩与中心极限定理

STEP II 1996，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/96-s2-q14

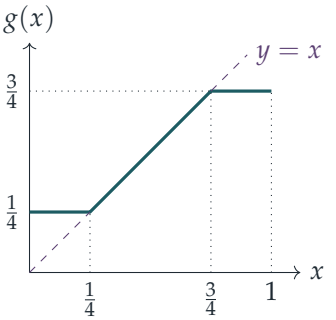
关键思路

Y 就是把 X 的两端“截平”：所有低于 $\frac{1}{4}$ 的概率都堆在 $\frac{1}{4}$ 这一个值上，所有高于 $\frac{3}{4}$ 的概率都堆在 $\frac{3}{4}$ 上。所以 Y 是一个混合型随机变量（mixed random variable），有两个点质量（point mass）和中间一段连续部分，而任何期望值（expectation）$\mathrm{E}(g(Y))$ 也同样分成三部分。最后一部分涉及大量独立复本之和，这就要用上中心极限定理（central limit theorem）。为此我们只需要 Y 的平均值和方差（variance），而前面几部分已经把这项工作做好了。

解答

Y 的矩

由于 $X \sim \mathrm{U}[0,1]$，我们有 $\mathrm{P}(X \le \frac{1}{4}) = \mathrm{P}(X \ge \frac{3}{4}) = \frac{1}{4}$，而在 $[\frac{1}{4}, \frac{3}{4}]$ 上 X 的概率密度（density）是 1。写成 $Y = g(X)$，其中 g 是下图所示的截平函数（clipping function），便得 $\mathrm{E}(Y^n) = \int_0^1 g(x)^n \,\mathrm{d}x$。



把积分在 $\frac{1}{4}$ 和 $\frac{3}{4}$ 处分开，

$$\begin{aligned}\mathrm{E}(Y^n) &= \left(\tfrac{1}{4}\right)^n \cdot \tfrac{1}{4} + \int_{1/4}^{3/4} x^n \,\mathrm{d}x + \left(\tfrac{3}{4}\right)^n \cdot \tfrac{1}{4} \\ &= \frac{1}{4^{n+1}} + \frac{3^{n+1}-1}{(n+1)4^{n+1}} + \frac{3^n}{4^{n+1}} \\ &= \frac{(n+1)(1+3^n)+3^{n+1}-1}{(n+1)4^{n+1}}.\end{aligned}$$

分子是 $n+1+(n+1)3^n+3\cdot 3^n-1 = n+(n+4)3^n$，所以

$$\boxed{\mathrm{E}(Y^n) = \frac{n+(n+4)\,3^n}{(n+1)\,4^{n+1}}} \qquad (n \geq 1).$$

与 X 比较

对于均匀分布（uniform distribution）的变量，$\mathrm{E}(X^n) = \int_0^1 x^n \,\mathrm{d}x = \dfrac{1}{n+1}$。

取 $n=1$，公式给出 $\mathrm{E}(Y) = \dfrac{1+5\cdot 3}{2\cdot 16} = \dfrac{1}{2} = \mathrm{E}(X)$，正如所求。（这并不令人意外：截平把两端的质量向内移动了相同的幅度，而且关于 $\frac{1}{2}$ 对称。）

取 $n=2$，$\mathrm{E}(Y^2) = \dfrac{2+6\cdot 9}{3\cdot 64} = \dfrac{56}{192} = \dfrac{7}{24}$，而 $\mathrm{E}(X^2) = \frac{1}{3} = \frac{8}{24}$。因此

$$\mathrm{E}(X^2) - \mathrm{E}(Y^2) = \frac{8}{24} - \frac{7}{24} = \frac{1}{24},$$

正如所求。

对于一般的 $n \geq 2$，把两个矩都写成以 $(n+1)4^{n+1}$ 为公分母的形式：

$$\mathrm{E}(X^n) = \frac{4^{n+1}}{(n+1)4^{n+1}}.$$

所以我们要证明 $4^{n+1} > n+(n+4)3^n$。利用 $4^{n+1} = 4(3+1)^n$ 和二项式定理（binomial theorem），并把最高的两项分开写出，

$$4^{n+1} = 4\cdot 3^n + 4n\cdot 3^{n-1} + 4\sum_{k=0}^{n-2}\binom{n}{k}3^k.$$

减去 $n+(n+4)3^n = 4\cdot 3^n + 3n\cdot 3^{n-1} + n$，得

$$4^{n+1} - (n+(n+4)3^n) = n\,3^{n-1} - n + 4\sum_{k=0}^{n-2}\binom{n}{k}3^k.$$

? **想一想** 为甚么这个差是正数?

当 $n \geq 2$ 时，$3^{n-1} \geq 3$，所以 $n\,3^{n-1} - n \geq 2n > 0$。那个和也是正数，因为它包含 $k = 0$ 的一项，其值为 1，而且每一项都是非负的。因此这个差严格为正，于是

$$\mathrm{E}(X^n) > \mathrm{E}(Y^n) \qquad \text{对所有} n \geq 2.$$

（当 $n = 1$ 时，这个差是 $3^0 - 1 + 0 = 0$，与 $\mathrm{E}(X) = \mathrm{E}(Y)$ 吻合。）

240 000 个复本之和

由上一部分，

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \frac{7}{24} - \frac{1}{4} = \frac{1}{24}.$$

设 $S = Y_1 + \cdots + Y_N$，其中 $N = 240\,000$。由于各 Y_i 互相独立（independent），期望值和方差都可以相加：

$$\mathrm{E}(S) = \frac{N}{2} = 120\,000, \qquad \mathrm{Var}(S) = \frac{N}{24} = 10\,000,$$

所以 S 的标准差（standard deviation）是 100。各 Y_i 独立同分布（independent and identically distributed），方差有限，而 N 非常大，所以根据中心极限定理，S 近似服从 $\mathrm{N}(120\,000,\, 100^2)$。因此

$$\mathrm{P}(S < K) \approx \Phi\Big(\frac{K - 120\,000}{100}\Big) = \frac{3}{4} \quad \Longrightarrow \quad \frac{K - 120\,000}{100} = \Phi^{-1}(0.75) \approx 0.6745,$$

所以

$$K \approx 120\,000 + 67.45 \approx 120\,067.$$

这里不需要连续性校正（continuity correction）：Y 并非只取整数值，而且标准差为 100，任何这类校正对这个准确度根本没有影响。

验算

- $\mathrm{E}(Y^n)$ 的公式所给的值介乎 $(\frac{1}{4})^n$ 与 $(\frac{3}{4})^n$ 之间，这是必然的，因为 $\frac{1}{4} \leq Y \leq \frac{3}{4}$。当 $n = 3$ 时，公式给出 $\frac{3+7\cdot 27}{4\cdot 256} = \frac{192}{1024} = \frac{3}{16} = 0.1875$，而用中点法则（midpoint rule）计算 $g(x)^3$ 的积分，结果吻合至小数点后十位。
- $\mathrm{Var}(Y) = \frac{1}{24} < \frac{1}{12} = \mathrm{Var}(X)$：截平把数值拉向平均值，所以离散程度理应下降。
- 模拟 20 000 个由 2400 个 Y 的复本相加而成的和，得出标准化后的上四分位数（upper quartile）约为 0.678，接近 0.6745，这支持了正态近似（normal approximation）。

常见错误

- 把 Y 当作在 $[\frac{1}{4}, \frac{3}{4}]$ 上均匀分布，因而忘记了两端各有大小为 $\frac{1}{4}$ 的点质量。在该区间上均匀分布的变量，密度是 2，不是 1。

- 用了 $\operatorname{Var}(S) = N^2 \operatorname{Var}(Y)$ 而不是 $N \operatorname{Var}(Y)$。独立性使方差可以相加，所以标准差按 $\sqrt{N}$ 的方式增长。
- 在最后一步弄错正负号。$\mathrm{P}(S < K) = \frac{3}{4} > \frac{1}{2}$，所以 K 必定高于平均值。

STEP 之外

Y 是 X 经截平（又称“缩尾”, Winsorised）后的版本, 它的分布既不是离散的, 也不是连续的。测度论概率（measure-theoretic probability）通过分布函数（distribution function）统一处理这类分布，每个期望值都变成单一个积分 $\int g \, \mathrm{d}F$。在平均值相等的情况下，不等式 $\mathrm{E}(X^n) > \mathrm{E}(Y^n)$ 是凸序（convex ordering）的一个例子：X 是 Y 的“保均值扩散”（mean-preserving spread），所以对每个凸函数 ϕ 都有 $\mathrm{E}(\phi(X)) \geq \mathrm{E}(\phi(Y))$。最后用到的中心极限定理，其证明见 Grimmett and Welsh, *Probability: An Introduction*（一本概率论入门教科书）。

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。