详细解答

STEP II 2023，第 12 题

概率与统计 · 最大值的分布

STEP II 2023，第 12 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 2023 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

www.ocr.org.uk/Images/707102-2023-past-papers.zip

关键思路

若干个数的最大值不超过 t，当且仅当其中每一个都不超过 t，因此 Y 的累积分布函数（cumulative distribution function）是单个 X_i 的累积分布函数的幂。其余一切都源于使用累积分布函数而不是密度函数：中位数（median）可直接解方程 $\mathrm{P}(Y \le m) = \frac{1}{2}$ 得出，而均值（mean）则可通过分部积分（integration by parts）简洁地求得，使积分中出现的是分布函数而非密度函数。

解答

第 (i) 部分

记 $F(t) = \mathrm{P}(X_1 \le t)$。当 $0 \le t \le \pi$ 时，

$$F(t) = \int_0^t \tfrac{1}{2}\sin x \,\mathrm{d}x = \tfrac{1}{2}(1 - \cos t).$$

事件 $\{Y \le t\}$ 发生，当且仅当 $X_1 \le t$、$X_2 \le t$、…、$X_n \le t$ 全部发生：若最大者不超过 t，则其余各个亦不超过 t，反之亦然。各 X_i 独立同分布（independent and identically distributed），故这个交集的概率等于各个概率之积：

$$\mathrm{P}(Y \le t) = \mathrm{P}(X_1 \le t)\,\mathrm{P}(X_2 \le t) \cdots \mathrm{P}(X_n \le t) = [\mathrm{P}(X_1 \le t)]^n.$$

因此，以 G 记 Y 的累积分布函数，

$$G(t) = \frac{(1 - \cos t)^n}{2^n} \qquad (0 \le t \le \pi),$$

且当 $t<0$ 时 $G(t)=0$，当 $t>\pi$ 时 $G(t)=1$。用链式法则（chain rule）求导，得 Y 的概率密度函数（probability density function）为

$$g(t)=\frac{n\,(1-\cos t)^{n-1}\sin t}{2^n}\quad \text{当} 0\le t\le \pi,$$

其他情况下为零。

第 (ii) 部分

中位数 $m=m(n)$ 满足 $G(m)=\frac{1}{2}$：

$$\left(\frac{1-\cos m}{2}\right)^n=\frac{1}{2}\iff \frac{1-\cos m}{2}=2^{-1/n}$$
$$\iff \cos m=1-2^{1-1/n},$$

由于 $(1-\cos m)/2\ge 0$，这里取正实数 n 次方根。由于 m 必须位于 $[0,\pi]$ 内，而余弦函数在此区间上是一一对应（one-to-one）的，

$$m(n)=\arccos(1-2^{1-1/n}).$$

当 n 增大时，$1/n$ 减小，故 $2^{1-1/n}$ 增大（趋向 2），而 $1-2^{1-1/n}$ 减小（趋向 -1）。由于 arccos 是递减函数，$m(n)$ 随 n 递增，且当 $n\to\infty$ 时 $m(n)\to\arccos(-1)=\pi$。（当 $n=1$ 时得 $m(1)=\arccos 0=\pi/2$，这正符合 $\frac{1}{2}\sin x$ 关于 $\pi/2$ 的对称性。）这是合理的：取越来越多个值的最大值，结果会被推向取值范围的顶端。

第 (iii) 部分

按定义，$\mu(n)=\int_0^\pi t\,g(t)\,\mathrm{d}t$。由于 $g=G'$，取 $u=t$ 及 $\mathrm{d}v=g(t)\,\mathrm{d}t$、$v=G(t)$ 进行分部积分：

$$\mu(n)=\Big[t\,G(t)\Big]_0^\pi-\int_0^\pi G(t)\,\mathrm{d}t$$
$$=\pi\cdot 1-0-\frac{1}{2^n}\int_0^\pi(1-\cos x)^n\,\mathrm{d}x,$$

其中用到 $G(\pi)=(\frac{1+1}{2})^n=1$。这就是所求的公式。

> ? **想一想** 为甚么要用分部积分，而不是尝试直接计算 $\int t\,(1-\cos t)^{n-1}\sin t\,\mathrm{d}t$?

目标公式含有 $(1-\cos x)^n$，即 $2^nG(x)$，是分布函数而非密度函数。分部积分正是把 $t\,g(t)$ 换成 $G(t)$ 的一步，同时亦消去了麻烦的因子 t。

(a) 设 $h(x)=\frac{1}{2}(1-\cos x)$，则在 $[0,\pi]$ 上 $0\le h(x)\le 1$，且当 $0<x<\pi$ 时 $0<h(x)<1$。公式变为 $\mu(n)=\pi-\int_0^\pi h(x)^n\,\mathrm{d}x$，故

$$\mu(n+1)-\mu(n)=\int_0^\pi\left(h(x)^n-h(x)^{n+1}\right)\mathrm{d}x=\int_0^\pi h(x)^n(1-h(x))\,\mathrm{d}x.$$

在 $0 < x < \pi$ 上，被积函数连续且严格为正，因为在该处 $h(x) > 0$ 且 $1 - h(x) > 0$。所以积分严格为正，对每个 $n \geq 1$ 都有 $\mu(n+1) > \mu(n)$：$\mu(n)$ 随 n 递增。

(b) 取 $n = 2$，展开并利用 $\cos^2 x = \frac{1}{2}(1 + \cos 2x)$：

$$\int_0^{\pi} (1 - \cos x)^2 \, \mathrm{d}x = \int_0^{\pi} \left(1 - 2\cos x + \tfrac{1}{2} + \tfrac{1}{2}\cos 2x\right) \mathrm{d}x = \tfrac{3}{2}\pi,$$

因为 $\cos x$ 和 $\cos 2x$ 在 $[0, \pi]$ 上的积分均为零。因此

$$\mu(2) = \pi - \tfrac{1}{4} \cdot \tfrac{3}{2}\pi = \frac{5\pi}{8}.$$

由第 (ii) 部分，$m(2) = \arccos(1 - \sqrt{2})$。$5\pi/8$ 与 $m(2)$ 均位于 $[0, \pi]$ 内，而余弦函数在此区间上严格递减，故

$$\mu(2) < m(2) \iff \cos \tfrac{5\pi}{8} > \cos m(2) = 1 - \sqrt{2}.$$

? **想一想** 怎样求得 $\cos \frac{5\pi}{8}$ 的准确值?

写成 $\cos \frac{5\pi}{8} = \cos(\frac{\pi}{2} + \frac{\pi}{8}) = -\sin \frac{\pi}{8}$，并利用半角公式（half-angle formula）$\sin^2 \frac{\pi}{8} = \frac{1}{2}(1 - \cos \frac{\pi}{4}) = \frac{2-\sqrt{2}}{4}$。由于 $\sin \frac{\pi}{8} > 0$，

$$\cos \tfrac{5\pi}{8} = -\tfrac{1}{2}\sqrt{2 - \sqrt{2}}.$$

所以我们需要证明 $-\frac{1}{2}\sqrt{2-\sqrt{2}} > 1 - \sqrt{2}$，即 $\sqrt{2} - 1 > \frac{1}{2}\sqrt{2-\sqrt{2}}$。两边均为正数，故此式等价于两边平方后的不等式，再乘以 4：

$$\begin{aligned} 4(3 - 2\sqrt{2}) > 2 - \sqrt{2} &\iff 10 > 7\sqrt{2} \\ &\iff 100 > 98, \end{aligned}$$

其中最后一步是把两个正数平方。此式成立，而每一步都是等价变换，故 $\cos \frac{5\pi}{8} > 1 - \sqrt{2}$，因此 $\mu(2) < m(2)$。

验算

- 当 $n = 1$ 时，$Y = X_1$ 关于 $\pi/2$ 对称，而确实 $\mu(1) = \pi - \frac{1}{2}\int_0^{\pi}(1 - \cos x)\,\mathrm{d}x = \pi/2 = m(1)$。
- 数值上，$\mu(2) = 5\pi/8 \approx 1.9635$，$m(2) = \arccos(1 - \sqrt{2}) \approx 1.9979$，所以均值略低于中位数。对 n 由 1 至 7 的每个值各以电脑模拟 10^5 个最大值，结果与两条公式的误差均在 0.01 以内（例如 $\mu(3) \approx 2.1598$，$m(3) \approx 2.1986$），而且两个数列都递增。
- 当 $n = 2$ 时，Y 的密度函数为 $\frac{1}{2}(1 - \cos t)\sin t$，属左偏（skewed to the left）分布（有一条伸向 0 的长尾），而左偏分布的均值通常低于中位数。

常见错误

- 忘记论证 $\mathrm{P}(Y \le t) = [F(t)]^n$：既需要“全部都不超过 t”这个等价关系，也需要独立性。
- 做了分部积分，却忘记边界项 $\left[tG(t)\right]_0^{\pi} = \pi$，而开首的 π 正是由此而来。
- 通过余弦比较角的大小时，没有指出余弦函数在 $[0, \pi]$ 上是*递减*的，结果把不等号方向弄反。

STEP 之外

样本的最大值是*顺序统计量*（order statistic）的一个例子，而第 (i) 部分的技巧可以求出任何样本最大值的分布。恒等式 $\mathrm{E}[Y] = \pi - \int_0^{\pi} G(t)\,\mathrm{d}t$ 是非负随机变量的一般尾部积分公式（tail-integral formula）$\mathrm{E}[Y] = \int_0^{\infty} \mathrm{P}(Y > t)\,\mathrm{d}t$ 的特例，你在大学概率课程的初期就会遇到。这两个概念在 Grimmett 与 Welsh 的 *Probability: An Introduction*（《概率导论》）中均有深入讨论。

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。