详细解答

STEP II 2000，第 14 题

概率与统计 • 中位数的分布

STEP II 2000，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/00-s2-q14

本题在 *Beyond STEP* 第 13.4 节平均数还是中位数？中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch13/mean-or-median.html

关键思路

中位数（median）的概率密度函数（probability density function）是形如 $y^r(1-y)^s$ 的多项式，所以题目给出的积分已包办所有工作：先用一次求归一化常数（normalising constant），再把 y 的幂次分别提高一和提高二，求出首两个矩（moments）。至于最后一部分，Y 和 $\bar{X}$ 都以 $\frac{1}{2}$ 为中心，而对正态分布（normal distribution）来说，落在距离中心某固定距离之内的概率只取决于标准差（standard deviation）。离散程度较小的那个变量胜出。

解答

常数 k

密度函数的积分必须等于 1。在题目给出的结果中取 $r = s = n$，

$$1 = \int_0^1 ky^n(1-y)^n \,\mathrm{d}y = k\,\frac{n!\,n!}{(2n+1)!}, \qquad \text{所以} \qquad k = \frac{(2n+1)!}{(n!)^2},$$

正如所求。

平均值

密度函数关于 $y=\frac{1}{2}$ 对称（把 y 换成 $1-y$，函数不变），所以 $\mathrm{E}(Y)=\frac{1}{2}$。要用积分来确认这一点，取 $r=n+1$、$s=n$：

$$\mathrm{E}(Y)=k\int_0^1 y^{n+1}(1-y)^n\,\mathrm{d}y=\frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+1)!\,n!}{(2n+2)!}=\frac{n+1}{2n+2}=\frac{1}{2}.$$

方差

现在取 $r=n+2$、$s=n$：

$$\mathrm{E}(Y^2)=\frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!}=\frac{(n+2)(n+1)}{(2n+3)(2n+2)}=\frac{n+2}{2(2n+3)}.$$

因此

$$\mathrm{Var}(Y)=\frac{n+2}{2(2n+3)}-\frac{1}{4}=\frac{2(n+2)-(2n+3)}{4(2n+3)}=\frac{1}{4(2n+3)}.$$

比较中位数与平均值

每个 X_i 都在 $[0,1]$ 上服从均匀分布（uniform distribution），所以 $\mathrm{E}(X_i)=\frac{1}{2}$，而 $\mathrm{Var}(X_i)=\frac{1}{12}$。由于各 X_i 互相独立（independent），

$$\mathrm{E}(\bar{X})=\tfrac{1}{2},\qquad \mathrm{Var}(\bar{X})=\frac{1}{(2n+1)^2}\cdot(2n+1)\cdot\frac{1}{12}=\frac{1}{12(2n+1)}.$$

想一想 两个变量都以 $\frac{1}{2}$ 为中心。哪一个数值决定了谁较有可能落在距离中心 $d/\sqrt{n}$ 之内?

答案是标准差。当 n 很大时，我们可以取 $Y\sim\mathrm{N}(\frac{1}{2},\sigma_Y^2)$ 和 $\bar{X}\sim\mathrm{N}(\frac{1}{2},\sigma_{\bar{X}}^2)$。若 $W\sim\mathrm{N}(\frac{1}{2},\sigma^2)$ 且 $c>0$，则以 Φ 表示标准正态分布函数（standard normal distribution function），有

$$\mathrm{P}\big(|W-\tfrac{1}{2}|<c\big)=\mathrm{P}\Big(|Z|<\frac{c}{\sigma}\Big)=2\Phi\Big(\frac{c}{\sigma}\Big)-1,$$

其中 $Z\sim\mathrm{N}(0,1)$。这是 σ 的严格递减函数，因为 Φ 严格递增。所以取 $c=d/\sqrt{n}$ 时，所求的不等式成立，当且仅当 $\sigma_Y>\sigma_{\bar{X}}$，即 $\mathrm{Var}(Y)>\mathrm{Var}(\bar{X})$：

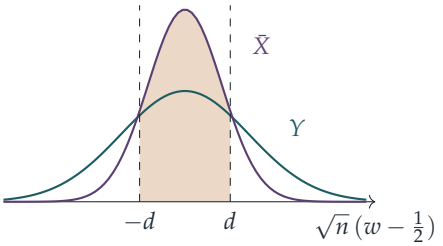
$$\begin{aligned}\frac{1}{4(2n+3)}>\frac{1}{12(2n+1)} &\iff 12(2n+1)>4(2n+3)\\ &\iff 24n+12>8n+12\\ &\iff n>0.\end{aligned}$$

这对每个 $n\geq 1$ 都成立。所以只要 n 大得足以令正态近似（normal approximation）有效，则不论正数 d 是甚么，Y 的概率都严格小于 $\bar{X}$ 的概率。证明至此完成。

要看看差距有多大，可以把变量标准化：

$$\frac{d/\sqrt{n}}{\sigma_Y} = 2d\sqrt{\frac{2n+3}{n}} \to 2\sqrt{2}\,d, \qquad \frac{d/\sqrt{n}}{\sigma_X} = d\sqrt{\frac{12(2n+1)}{n}} \to 2\sqrt{6}\,d.$$

渐近而言，平均值的标准差是中位数的 $\sqrt{3}$ 分之一。



图中以 $\sqrt{n}\,(w-\frac{1}{2})$ 为标度，画出两个极限密度函数。较窄的曲线（平均值）在任何一条中央带内都有较大的概率。

验算

- $n=0$: 一个均匀变量的“中位数”就是 X_1 本身, 而我们的公式给出 $k=1$ 和 $\mathrm{Var}(Y)=\frac{1}{12}$，正是均匀分布的方差，与预期相符。
- 当 $n=0$ 时两个方差都等于 $\frac{1}{12}$，而当 $n\geq 1$ 时两者便分开，这与上面的不等式 $n>0$ 吻合。
- 以 $n=50$（即 101 个均匀变量）进行模拟, 得到 $\mathrm{Var}(Y)\approx 0.00243$, 而 $1/412\approx 0.00243$。取 $d=0.3$ 时，模拟得到 $\mathrm{P}(|Y-\frac{1}{2}|<d/\sqrt{n})\approx 0.607$ 和 $\mathrm{P}(|\bar{X}-\frac{1}{2}|<d/\sqrt{n})\approx 0.859$, 与正态分布给出的数值 0.611 和 0.860 相当接近。

常见错误

- 写成 $\mathrm{Var}(\bar{X})=\frac{1}{12}$，或者只除以 $2n+1$ 一次，而不是先在分母得到 $(2n+1)^2$，再乘以 $2n+1$。
- 弄错方向：方差较大，落在中心附近的概率反而较小。
- 忘记了计算 $\mathrm{E}(Y^2)$ 时用的是 $r=n+2$，而不是 $r=2n$。

STEP 之外

中位数的密度函数是一个贝塔分布（Beta distribution）$\mathrm{Beta}(n+1,n+1)$，而题目给出的积分就是贝塔函数（Beta function）。你刚才所作的比较，正是渐近相对效率（asymptotic

relative efficiency）的概念：对均匀分布的数据，用样本平均值（sample mean）估计中心，其方差只有中位数的三分之一；但对厚尾（heavy-tailed）的数据，中位数却可以胜出。顺序统计量（order statistics）和贝塔函数在 Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）中有很好的讲解。

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。