详细解答

STEP III 2009，第 13 题

概率与统计 • 中心极限定理

STEP III 2009，第 13 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/09-s3-q13

关键思路

一切都是单一角度 θ 的函数：$X = \cos\theta$ 及 $Y = \sin\theta$。因此，X 的分布可由累积分布函数（cumulative distribution function）求得（找出圆上哪一段弧使 $\cos\theta \le x$），而平均值、方差（variance）和协方差（covariance）不过是三角函数在一整圈上的平均。在第 (ii) 部分，$\overline{X}$ 是许多独立同分布（independent, identically distributed）变量的平均，这正是中心极限定理（central limit theorem）适用的情况。

解答

第 (i) 部分

由于 P 在单位圆上，$X = \cos\theta$ 及 $Y = \sin\theta$，其中 θ 在 $[0, 2\pi)$ 上的密度为 $\frac{1}{2\pi}$。

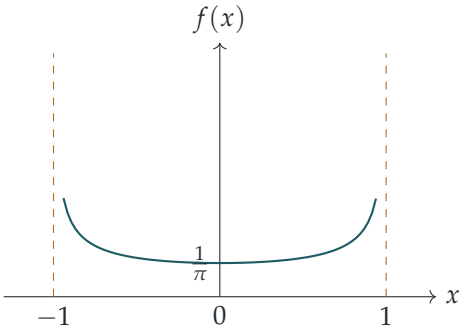
X 的密度。 显然 $-1 \le X \le 1$。取 $-1 < x < 1$，并设 $\alpha = \arccos x \in (0, \pi)$。在 $[0, 2\pi)$ 上，$\cos\theta > x$ 当且仅当 $\theta < \alpha$ 或 $\theta > 2\pi - \alpha$，这是一段总长度为 2α 的弧。因此

$$F(x) = \mathrm{P}(X \le x) = 1 - \frac{2\alpha}{2\pi} = 1 - \frac{\arccos x}{\pi}.$$

利用 $\dfrac{\mathrm{d}}{\mathrm{d}x}\arccos x = -\dfrac{1}{\sqrt{1-x^2}}$ 求导，得

$$f(x) = \frac{1}{\pi\sqrt{1-x^2}} \qquad (-1 < x < 1),$$

而在其他情况下 $f(x) = 0$。（简单验算一下：$\int_{-1}^{1} f(x)\,\mathrm{d}x = \frac{1}{\pi}[\arcsin x]_{-1}^{1} = 1$。）



密度在 $x = 0$ 处最小，在 $x = \pm 1$ 处趋向无穷大：点 P 在水平直径两端附近的水平移动最慢，所以 X 大部分时间都停留在那里。

方差。 由对称性（f 是偶函数），$\mathrm{E}(X) = 0$。直接用 θ 计算，

$$\mathrm{E}(X^2) = \frac{1}{2\pi}\int_0^{2\pi} \cos^2\theta \,\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi} \tfrac{1}{2}(1+\cos 2\theta)\,\mathrm{d}\theta = \frac{1}{2},$$

所以

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \tfrac{1}{2}.$$

同理（或因为 $\theta - \frac{\pi}{2}$ 同样在一整圈上均匀分布），$\mathrm{E}(Y) = 0$ 及 $\mathrm{Var}(Y) = \frac{1}{2}$。

不相关（uncorrelated）。 我们有

$$\mathrm{E}(XY) = \frac{1}{2\pi}\int_0^{2\pi} \cos\theta \sin\theta \,\mathrm{d}\theta = \frac{1}{4\pi}\int_0^{2\pi} \sin 2\theta \,\mathrm{d}\theta = 0,$$

所以 $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = 0 - 0 = 0$。因此 X 与 Y 不相关。

并非独立。 它们肯定不是独立的，因为 $X^2 + Y^2 = 1$：一旦知道 X, Y 就只能是 $\pm\sqrt{1-X^2}$ 其中之一。举一个具体事件为例：$\mathrm{P}(X > 0.8) > 0$ 及 $\mathrm{P}(Y > 0.8) > 0$，但 $\mathrm{P}(X > 0.8 \text{ 及 } Y > 0.8) = 0$，因为那样的话 $X^2 + Y^2 > 1.28$。若两者独立，这个联合概率便须等于那个（正的）乘积。相关系数为零只表示没有线性关系；这里的关系是一个圆，它足够对称，以致完全没有线性趋势。

第 (ii) 部分

每一对 (X_i, Y_i) 都服从第 (i) 部分的分布，而由于各点是独立选取的，下标不同的各对互相独立。

$\overline{X}$ 与 $\overline{Y}$ 不相关。 协方差是双线性（bilinear）的，所以

$$\mathrm{Cov}(\overline{X},\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i,Y_j).$$

当 $i \neq j$ 时，X_i 只取决于 P_i，Y_j 只取决于 P_j，而这两点互相独立，所以 $\mathrm{Cov}(X_i,Y_j)=0$。当 $i=j$ 时，由第 (i) 部分得 $\mathrm{Cov}(X_i,Y_i)=0$。每一项都等于零，所以 $\mathrm{Cov}(\overline{X},\overline{Y})=0$：$\overline{X}$ 与 $\overline{Y}$ 不相关。

近似概率。 各 X_i 独立同分布，平均值为 0，方差为 $\frac{1}{2}$。所以

$$\mathrm{E}(\overline{X})=0, \qquad \mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot\frac{n}{2} = \frac{1}{2n},$$

这里可以把方差相加，是因为各 X_i 互相独立。

> **想一想** X 一点也不像正态分布（normal distribution）（看看它的密度）。为甚么我们可以用正态近似?

中心极限定理只要求各项独立同分布且方差有限，不论其分布形状如何。所以当 n 很大时，$\overline{X}$ 近似服从 $\mathrm{N}(0,\frac{1}{2n})$，标准差（standard deviation）为 $\sigma = \dfrac{1}{\sqrt{2n}}$。现在

$$\sqrt{\frac{2}{n}} = \frac{2}{\sqrt{2n}} = 2\sigma,$$

所以设 $Z\sim\mathrm{N}(0,1)$，则

$$\begin{aligned}\mathrm{P}\left(|\overline{X}|\le\sqrt{\tfrac{2}{n}}\right) = \mathrm{P}\left(\frac{|\overline{X}|}{\sigma}\le 2\right) &\approx \mathrm{P}(|Z|\le 2)\\ &= 2\Phi(2)-1 = 2(0.9772)-1 \approx 0.954 \approx 0.95,\end{aligned}$$

正如题目所求。

验算

- 方差 $\frac{1}{2}$ 是合理的：$X^2+Y^2=1$，而由对称性 $\mathrm{E}(X^2)=\mathrm{E}(Y^2)$，所以两者都必定是 $\frac{1}{2}$。根本不用积分!
- 用 4×10^5 个随机角度进行模拟，结果在多个 x 值上与 $F(x)=1-\frac{1}{\pi}\arccos x$ 吻合至约小数点后三位，并得出 $\mathrm{Var}(X)\approx0.5007$ 及 $\mathrm{Cov}(X,Y)\approx0.001$。
- 取 $n=50$ 并重复 40 000 次，模拟所得的 $\mathrm{Var}(\overline{X})$ 为 0.00997（理论值 0.01），而 $|\overline{X}|\le\sqrt{2/50}=0.2$ 所占的比例为 0.954，对比 $\mathrm{P}(|Z|\le2)=0.9545$。

常见错误

- 对 $1 - \frac{1}{\pi}\arccos x$ 求导时弄错正负号，得出一个负的“密度”。密度永远不可能是负数，所以要检查正负号。
- 由协方差为零推断两者独立。独立可推出协方差为零，但反过来并不成立。
- 使用 $\mathrm{Var}(\overline{X}) = \frac{1}{2}$ 或 $\frac{1}{2} \cdot \frac{1}{\sqrt{n}}$：$n$ 个独立副本的平均值，其方差是单个方差除以 n；按 $1/\sqrt{n}$ 的比例变化的是*标准差*。

STEP 之外

密度 $1/(\pi\sqrt{1-x^2})$ 就是*反正弦分布*（arcsine distribution），它出现的频繁程度出人意表，例如它是随机游走（random walk）处于正值的时间所占比例的极限分布。这道题的其余部分，谈的是不相关与独立之间的差距，以及中心极限定理；在大学一年级的概率课程中，你会用矩母函数（moment generating function）或特征函数（characteristic function）来证明这个定理。事实上，$(\overline{X}, \overline{Y})$ 近似服从二元正态分布（bivariate normal distribution），而对于联合正态的变量，相关系数为零*确实*可推出独立。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率入门教材）对以上各点都有很好的讲解。

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。