详细解答

STEP III 1989，第 15 题

概率与统计 · 中位数的分布

> **STEP III 1989，第 15 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/89-s3-q15

关键思路

$2n+1$ 个数值的中位数（median）落在 z 附近，当且仅当其中一个数值落在 z 处的一个极小区间内，另有 n 个数值在它之下、n 个在它之上。第 (i) 及第 (ii) 部分给出了“全部在下”和“全部在上”这两部分的概率，再数一数有多少种方法选定哪些数值担当哪个角色，便得到密度。之后，利用密度的积分等于 1 这个事实，不费吹灰之力就能求出该积分；把同一技巧用于以 $n+1$ 代替 n 的情况，便得到方差。

解答

在整个解答中，X 在 $[-c,c]$ 上的密度为 $\frac{1}{2c}$，所以对于 $-c \le k \le c$，

$$\mathrm{P}(X > k) = \frac{c-k}{2c}, \qquad \mathrm{P}(X < k) = \frac{c+k}{2c}.$$

第 (i) 部分

这 n 个数值互相独立（independent），所以把概率相乘：

$$\mathrm{P}(\text{全部 } n \text{ 个数值均大于} k) = \left(\frac{c-k}{2c}\right)^n.$$

第 (ii) 部分

同理，

$$\mathrm{P}(\text{全部 } n \text{ 个数值均小于} k) = \left(\frac{c+k}{2c}\right)^n.$$

中位数的密度

设 $X_1, \ldots, X_{2n+1}$ 为样本，Z 为其中位数，即第 $(n+1)$ 小的数值。由于 X 是连续的，出现相同数值的概率为 0，所以我们不予理会。每个 X_i 都在 $[-c, c]$ 内，所以 Z 也在其中。

取 $(-c, c)$ 内的 z 和一个小的 $\delta z > 0$。略去 $(\delta z)^2$ 阶的项后，事件 $z < Z \leq z + \delta z$ 发生，当且仅当

- 其中一个数值落在 $(z, z+\delta z]$ 内，
- 其余数值中有 n 个小于 z，并且
- 余下的 n 个数值大于 $z + \delta z$。

（有两个或以上数值落在这个小区间内的情况，其概率为 $(\delta z)^2$ 阶，在取极限时无关重要。）

? **想一想** 把这 $2n+1$ 个数值分配到这三个角色，共有多少种方法?

落在小区间内的数值有 $2n+1$ 种选择，然后从其余数值中选出 n 个放在 z 之下，有 $\binom{2n}{n}$ 种方法。于是共有

$$(2n+1)\binom{2n}{n} = \frac{(2n+1)!}{(n!)^2}$$

种分配方法。这些分配互斥（mutually exclusive），而且每一种的概率相同。对于一种固定的分配，由独立性以及第 (i) 和第 (ii) 部分（取 $k = z$），其概率为

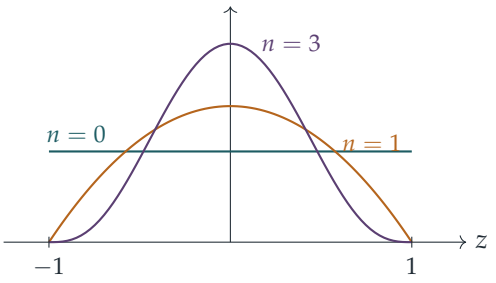
$$\frac{\delta z}{2c}\left(\frac{c+z}{2c}\right)^n \left(\frac{c-z-\delta z}{2c}\right)^n.$$

对所有分配求和，除以 δz，再令 $\delta z \to 0$，便得 Z 的密度为

$$\begin{aligned} g(z) &= \frac{(2n+1)!}{(n!)^2} \cdot \frac{1}{2c} \cdot \frac{(c+z)^n (c-z)^n}{(2c)^{2n}} \\ &= \frac{(2n+1)!}{(n!)^2 (2c)^{2n+1}} (c^2 - z^2)^n, \qquad -c \leq z \leq c, \end{aligned}$$

其中用到 $(c+z)(c-z) = c^2 - z^2$。这就是所要求的结果。

下图显示 $c = 1$ 时的密度。当 $n = 0$ 时，“中位数”只是单独一个数值，服从均匀分布（uniform distribution）。随着 n 增大，中位数愈来愈集中在 0 附近。



积分

概率密度（probability density）在其取值范围上的积分等于 1，所以

$$\frac{(2n+1)!}{(n!)^2(2c)^{2n+1}}\int_{-c}^{c}(c^2-z^2)^n\,\mathrm{d}z=1,$$

因此

$$\int_{-c}^{c}(c^2-z^2)^n\,\mathrm{d}z=\frac{(n!)^2(2c)^{2n+1}}{(2n+1)!}.$$

这对每个整数 $n\geq 0$ 都成立，稍后我们就会用到这一点。

平均值与方差

密度 g 是 z 的偶函数（even function），所以 $zg(z)$ 是奇函数，它在对称区间 $[-c,c]$ 上的积分为零：

$$\mathrm{E}(Z)=0.$$

因此 $\mathrm{var}(Z)=\mathrm{E}(Z^2)$。记 $K_n=\dfrac{(2n+1)!}{(n!)^2(2c)^{2n+1}}$ 及 $I_n=\displaystyle\int_{-c}^{c}(c^2-z^2)^n\,\mathrm{d}z=\frac{1}{K_n}$。

想一想 怎样才能不展开而求出 $z^2(c^2-z^2)^n$ 的积分?

写成 $z^2=c^2-(c^2-z^2)$。于是

$$\begin{aligned}\mathrm{E}(Z^2)&=K_n\int_{-c}^{c}\big(c^2-(c^2-z^2)\big)(c^2-z^2)^n\,\mathrm{d}z\\&=K_n\big(c^2I_n-I_{n+1}\big)=c^2-\frac{K_n}{K_{n+1}},\end{aligned}$$

其中用到 $K_nI_n=1$ 和 $I_{n+1}=1/K_{n+1}$，后者就是以 $n+1$ 代替 n 的积分结果。现在

$$\frac{K_n}{K_{n+1}}=\frac{(2n+1)!}{(2n+3)!}\cdot\frac{((n+1)!)^2}{(n!)^2}\cdot\frac{(2c)^{2n+3}}{(2c)^{2n+1}}=\frac{(n+1)^2\cdot 4c^2}{(2n+2)(2n+3)}=\frac{2(n+1)c^2}{2n+3}.$$

因此

$$\mathrm{var}(Z)=c^2\left(1-\frac{2n+2}{2n+3}\right)=\frac{c^2}{2n+3}.$$

验算

- 当 $n=0$ 时，中位数就是 X 的一个数值。密度变成 $\frac{1}{2c}$，积分为 $2c$，方差为 $\frac{c^2}{3}$，正是 $[-c,c]$ 上均匀分布的方差。
- 当 $n=1$ 及 $c=1$ 时，积分公式给出 $\int_{-1}^{1}(1-z^2)\,\mathrm{d}z=\frac{1\cdot 8}{6}=\frac{4}{3}$，结果正确。
- 以 10^5 个大小为 5 的样本（$n=2$）并取 $c=1.7$ 进行模拟，得到中位数的方差为 0.415，而 $c^2/7\approx 0.413$。

常见错误

- 忘记了选择中间数值所带来的因子 $2n+1$，只用了 $\binom{2n}{n}$。
- 漏掉中间数值本身的密度所带来的因子 $\frac{1}{2c}$，结果得到的幂是 $(2c)^{2n}$ 而不是 $(2c)^{2n+1}$。
- 用二项式展开 $(c^2-z^2)^n$ 来求 $\mathrm{E}(Z^2)$。这样做是可行的，但所得的和很难化简。

STEP 之外

中位数是样本的*顺序统计量*（order statistics）之一。同样的计数论证可以给出任何连续分布的 m 个数值中第 r 小者的密度，即 $\frac{m!}{(r-1)!\,(m-r)!}F^{r-1}(1-F)^{m-r}f$。对于均匀分布，这些都是*贝塔分布*（beta distributions），而你所推导出的积分正是贝塔函数 $B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ 的一个特例。方差 $c^2/(2n+3)$ 以 $1/n$ 的速度缩小，这是迈向样本中位数的渐近正态性（asymptotic normality）的第一步。Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）对顺序统计量和贝塔分布有很好的讲解。

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。