13.4 平均数还是中位数?

工作纸

题目

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

提示

请先自行尝试解题。如果遇到困难，请逐一阅读提示，每读完一个便回到题目再试。标有停一停，试一试的提示是关键步骤。每个提示之后均留有空位供你演算。

> **想一想** 为甚么中位数的密度（density）与 $y^n(1-y)^n$ 成正比？（要 Y 接近 y，$2n+1$ 个值之中须有多少个在其下，又有多少个大于 y？）

阶段 1 常数

涵盖：常数 k

> **想一想** 概率密度函数（density）的积分等于 1。利用题目给出的积分，把两个幂次都取为 n。

阶段 2 平均值与方差

涵盖：平均值（mean）与方差（variance）

想一想 这个密度函数是否关于 $\frac{1}{2}$ 对称（symmetric）？由此可得 $\mathrm{E}(Y)$。

停一停，试一试 要求 $\mathrm{E}(Y^2)$，可再次利用题目给出的积分，这次幂次取 $n+2$ 和 n，然后化简当中的阶乘（factorial）。

阶段 3　与平均值比较

涵盖：比较中位数（median）与平均值

想一想　$\bar{X}$ 的平均值和方差分别是甚么？比较 n 很大时这两个方差的大小。

停一停，试一试　把 Y 和 $\bar{X}$ 都视为平均值为 $\frac{1}{2}$ 的正态分布（normal distribution）。用 Φ 写出各个概率。哪一个较大？为甚么？

背后的数学

效率

两个无偏估计量（unbiased estimator）的*相对效率*（relative efficiency）是它们方差之比；它说明其中一个需要多少倍的观测值，才能与另一个看齐。对于均匀分布的数据，中位数相对于平均值的效率是 $\frac{1}{3}$。对于正态分布的数据，这个数字是 $\frac{2}{\pi} \approx 0.64$：中位数大约需要多 57% 的观测值。但遇上厚尾（heavy-tailed）数据，形势就逆转了。对于柯西（Cauchy）分布（Section 12.22），平均值毫无用处，因为 n 个观测值的平均并不比单独一个观测值好，而中位数却表现良好。对于拉普拉斯（Laplace）分布（Section 12.24），中位数的效率是平均值的两倍。

稳健统计学

中位数是*稳健*（robust）的：即使把接近一半的数据改成任意数值，它也不会移动太远。平均值却没有这种保障：一个错误的数值就足以把它毁掉。在 1960 及 1970 年代，John Tukey、Peter Huber 等人发展出*稳健统计学*（robust statistics），即一些估计量，对干净的数据而言效率几乎与平均值一样高，却又能抵御离群值（outlier），例如*截尾平均值*（trimmed mean）（舍去最大和最小的 10%，再取余下数据的平均），跳水、体操等运动项目的评分就采用这种方法。

拉普拉斯与中位数

拉普拉斯在 1774 年证明，假如误差服从双指数（拉普拉斯）分布，则真实值的最佳估计是中位数；高斯（Gauss）在 1809 年证明，假如误差服从正态分布，则最佳估计是平均值，他并由此得出最小二乘法（method of least squares）。谁是谁非，取决于误差的表现如何；每遇上一种新的数据，统计学家至今仍会为这个问题争论一番。

延伸练习

1 **中位数的密度。**用数算的方法推导密度 $\frac{(2n+1)!}{(n!)^2}y^n(1-y)^n$：一个值落在 $[y, y+\delta y]$ 内，n 个在其下，n 个在其上。

2 **三个观测值。**当 $n=1$（三个均匀观测值）时，求 $\mathrm{Var}(Y)$ 和 $\mathrm{Var}(\bar{X})$ 的准确值。哪一个较小?

3 **中列数。**中心的另一个估计量是中列数（midrange）$\frac{1}{2}(\max+\min)$。证明对于均匀分布的数据，若有 $m=2n+1$ 个观测值，它的方差是 $\frac{1}{2(m+1)(m+2)}$。它与平均值相比如何?（试与 Section 13.3 比较。）

4 **一个离群值。**十一个量度结果为 9.8、9.9、10.0、10.0、10.1、10.1、10.2、10.2、10.3、10.4 和 104（打字错误）。比较平均值和中位数。

延伸阅读

- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury），论及顺序统计量（order statistics）和效率。
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods***（Springer）。为非专家讲解稳健方法为何重要。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***（Harvard University Press）。

从历史角度看数据的汇总、平均值和中位数。

完整的引导式解答见第 13.4 节：

https://beyondstep.keithmcnulty.org/zh-hans/ch13/mean-or-median.html

每星期一道 STEP 题目，完整讲解：The STEP Up，以电子邮件免费发送，订阅网址：keithmcnulty.substack.com