详细解答

STEP III 2001，第 14 题

概率与统计 · 方差的无偏估计量

STEP III 2001，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/01-s3-q14

关键思路

整道题的关键只有一点：“样本的方差”所用的除数是 $n = 2$，因此会低估总体方差（population variance），而把它乘以二便可修正过来。只有两个值时，样本平均值（sample mean）就是两者的中点，所以每个值与它相距都是两值差距的一半。于是 V 只是单一个差 $X_1 - X_2$ 的函数，而最后要求的概率就变成正方形 $[0, a]^2$ 内的一块面积。

解答

X 的方差

X 的密度函数（density）在 $[0, a]$ 上是 $f(x) = \frac{1}{a}$。由对称性可知 $\mathrm{E}(X) = \frac{1}{2}a$，而

$$\mathrm{E}(X^2) = \int_0^a \frac{x^2}{a}\,\mathrm{d}x = \frac{a^2}{3}.$$

所以

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12},$$

正如题目所求。

样本方差

对于样本 X_1, X_2，样本平均值是 $\bar{X} = \frac{1}{2}(X_1 + X_2)$，而样本的方差就是各值与它的偏差平方的平均值：

$$V = \tfrac{1}{2}[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2].$$

由于 $X_1 - \bar{X} = \frac{1}{2}(X_1 - X_2)$ 及 $X_2 - \bar{X} = -\frac{1}{2}(X_1 - X_2)$，两个偏差平方都等于 $\frac{1}{4}(X_1 - X_2)^2$。因此

$$V = \tfrac{1}{2} \cdot 2 \cdot \tfrac{1}{4}(X_1 - X_2)^2 = \tfrac{1}{4}(X_1 - X_2)^2,$$

这正是题目给出的形式。

想一想 怎样不用二重积分（double integral）而求出 $\mathrm{E}((X_1 - X_2)^2)$?

由于 X_1 和 X_2 的分布相同，$\mathrm{E}(X_1 - X_2) = 0$，所以 $\mathrm{E}((X_1 - X_2)^2) = \mathrm{Var}(X_1 - X_2)$。因为 X_1 和 X_2 独立（independent），它们的方差可以相加（负号并无影响，因为 $\mathrm{Var}(-X_2) = \mathrm{Var}(X_2)$）：

$$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \frac{a^2}{12} + \frac{a^2}{12} = \frac{a^2}{6}.$$

因此

$$\mathrm{E}(2V) = \tfrac{1}{2}\,\mathrm{E}((X_1 - X_2)^2) = \tfrac{1}{2} \cdot \frac{a^2}{6} = \frac{a^2}{12} = \mathrm{Var}(X).$$

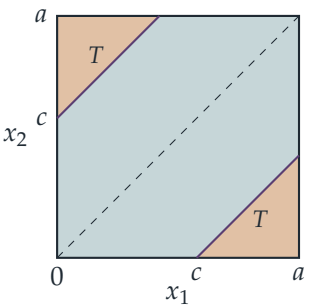
无论 a 取甚么值，$2V$ 的期望值（expected value）都等于它所估计的量，所以 $2V$ 是 X 的方差的无偏估计量（unbiased estimator）。（相比之下，$\mathrm{E}(V) = \frac{a^2}{24}$，只有真正方差的一半。）

概率

由于 $V \geq 0$ 而 $\frac{1}{12}a^2 > 0$，我们可以取平方根：

$$V < \frac{a^2}{12} \iff (X_1 - X_2)^2 < \frac{a^2}{3} \iff |X_1 - X_2| < \frac{a}{\sqrt{3}}.$$

因为 X_1 和 X_2 独立，而且各自服从均匀分布（uniform distribution），所以数对 (X_1, X_2) 均匀分布于正方形 $[0, a]^2$ 上。因此所求概率就是正方形面积中落在带状区域 $|x_1 - x_2| < c$ 内的比例，其中 $c = a/\sqrt{3} < a$。



余下的部分由角落两个着色的三角形 T 组成，每个都是等腰直角三角形，两条直角边的

长度均为 $a-c$。它们的总面积是 $2\cdot\frac{1}{2}(a-c)^2=(a-c)^2$，所以

$$\mathrm{P}\left(V<\frac{a^2}{12}\right)=1-\frac{(a-c)^2}{a^2}=1-\left(1-\frac{1}{\sqrt{3}}\right)^2=1-\left(\frac{4}{3}-\frac{2}{\sqrt{3}}\right)=\frac{2}{\sqrt{3}}-\frac{1}{3}.$$

也就是说，概率的准确值是

$$\mathrm{P}\left(V<\frac{a^2}{12}\right)=\frac{2\sqrt{3}-1}{3}.$$

取 $\sqrt{3}\approx 1.732$，这个值是 $\frac{2.464}{3}\approx 0.821$，所以准确至一位有效数字，概率是 0.8。

验算

- 答案与 a 无关，这是理所当然的：把 X 按某个倍数缩放，V 和 $\frac{1}{12}a^2$ 都会按该倍数的平方缩放。
- 取 $a=3$ 模拟 4×10^5 个数对，得出 $2V$ 的平均值为 0.751（目标值 0.75），而 $V<\frac{1}{12}a^2$ 所占的比例为 0.821，与 $\frac{1}{3}(2\sqrt{3}-1)=0.8214$ 吻合。
- 合理性检查：概率落在 $(0,1)$ 之内；若把 $\frac{1}{12}a^2$ 换成 $\frac{1}{4}a^2$（V 可取的最大值），则 $c=a$，而公式 $1-(1-c/a)^2$ 给出 1。

常见错误

- 计算“样本的方差”时用了除数 $n-1$：这样会得到 $\frac{1}{2}(X_1-X_2)^2$，而不是题目所述的 $\frac{1}{4}(X_1-X_2)^2$。题目要的是除数 n。
- 忘记了绝对值，结果只计算了两个三角形的其中一个（这样得出的概率会小于 $\frac{1}{2}$）。
- 写成 $\mathrm{Var}(X_1-X_2)=\mathrm{Var}(X_1)-\mathrm{Var}(X_2)=0$。

STEP 之外

这里的结果是一条一般定理在 $n=2$ 时的情况：从方差为 σ^2 的分布中抽取任何大小为 n 的样本，$\sum(X_i-\bar{X})^2$ 的平均值都是 $(n-1)\sigma^2$。这就是为甚么用除数 $n-1$（贝塞尔校正，Bessel's correction）会得到无偏估计量。大学一年级的统计学会进一步探讨：在各个无偏估计量之中，哪一个的方差最小。Grimmett 与 Welsh 的 *Probability: An Introduction*，以及 Blitzstein 与 Hwang 的 *Introduction to Probability*，都以本文所用的方式讲解和的期望值与方差。

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。