详细解答

STEP I 2005，第 13 题

概率与统计 • 对称分布

> **STEP I 2005，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/05-s1-q13

关键思路

题目除了给出两个累积概率（cumulative probability）和对称性（symmetry）之外，对分布的形状只字未提，所以一切都要由这些资料推导出来。关于 μ 对称，还附带送来两个事实：μ 的两侧各占一半概率，而且 $\mu - k\sigma$ 以下的尾部与 $\mu + k\sigma$ 以上的尾部概率相同。第 (ii) 部分的诀窍，是留意到题中提及的每一个容量（495、500、505 ml），对其中一种牛奶而言，都正好落在 μ、$\mu \pm \frac{1}{2}\sigma$ 或 $\mu + \sigma$ 的位置。

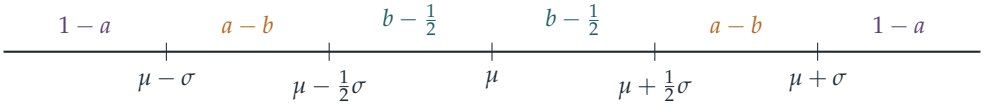
解答

全题把 X 视为连续（continuous）随机变量（它描述的是容量），因此单一数值的概率为零，写 $<$ 还是 $\leq$ 并无分别。

由关于 μ 的对称性，可得 $\mathrm{P}(X \leq \mu) = \frac{1}{2}$，并且对任意 k 都有 $\mathrm{P}(X \leq \mu - k\sigma) = \mathrm{P}(X \geq \mu + k\sigma)$。特别地，

$$\mathrm{P}(X \leq \mu - \tfrac{1}{2}\sigma) = 1 - b, \qquad \mathrm{P}(X \leq \mu - \sigma) = 1 - a.$$

不妨以 σ 为单位，从 μ 起量度，把每一段区间的概率记下来：



例如 $\mathrm{P}(\mu < X \leq \mu + \frac{1}{2}\sigma) = b - \frac{1}{2}$，而 $\mathrm{P}(\mu + \frac{1}{2}\sigma < X \leq \mu + \sigma) = a - b$；左边各段的概率则由对称性得出。

第 (i) 部分

把累积概率相减，

$$\mathrm{P}\big(\mu-\tfrac{1}{2}\sigma\le X\le\mu+\sigma\big)=\mathrm{P}(X\le\mu+\sigma)-\mathrm{P}\big(X<\mu-\tfrac{1}{2}\sigma\big)=a-(1-b)=a+b-1.$$

至于条件概率（conditional probability），分子所对应的事件是 $\{\mu-\frac{1}{2}\sigma\le X\le\mu+\frac{1}{2}\sigma\}$，其概率为 $b-(1-b)=2b-1$，而 $\mathrm{P}(X\ge\mu-\frac{1}{2}\sigma)=1-(1-b)=b$。因此

$$\mathrm{P}\big(X\le\mu+\tfrac{1}{2}\sigma\mid X\ge\mu-\tfrac{1}{2}\sigma\big)=\frac{2b-1}{b}.$$

第 (ii)(a) 部分

设 S 为该盒牛奶是脱脂奶的事件，F 为该盒牛奶是全脂奶的事件，则 $\mathrm{P}(S)=0.6$ 及 $\mathrm{P}(F)=0.4$。设 V 为以 ml 计的容量。

想一想 以 σ 为单位，500 ml 和 505 ml 在每种牛奶的分布中各落在甚么位置？

对脱脂奶（$\mu=500$, $\sigma=10$）而言，500 是 μ，505 是 $\mu+\frac{1}{2}\sigma$。对全脂奶（$\mu=495$, $\sigma=10$）而言，500 是 $\mu+\frac{1}{2}\sigma$，505 是 $\mu+\sigma$。从上图各段读出：

$$\mathrm{P}(V<505\mid S)=b,\qquad \mathrm{P}(500<V<505\mid S)=b-\tfrac{1}{2},$$
$$\mathrm{P}(V<505\mid F)=a,\qquad \mathrm{P}(500<V<505\mid F)=a-b.$$

根据全概率公式（law of total probability），

$$\mathrm{P}(V<505)=0.6b+0.4a,$$
$$\mathrm{P}(500<V<505)=0.6\big(b-\tfrac{1}{2}\big)+0.4(a-b)=0.4a+0.2b-0.3.$$

“多于 500 ml 而少于 505 ml”这个事件正是我们所需的交集，所以

$$\mathrm{P}(V>500\mid V<505)=\frac{0.4a+0.2b-0.3}{0.4a+0.6b}=\frac{4a+2b-3}{4a+6b}.$$

第 (ii)(b) 部分

随机抽取的牛奶盒数量非常大时，观察所得的比例可用来估计相应的概率。第一项资料表示 $\mathrm{P}(V\le505)=0.7$，按 (a) 部分的计算，即

$$0.6b+0.4a=0.7,\quad\text{即}\quad 4a+6b=7. \tag{1}$$

想一想 对每种牛奶而言，$\mathrm{P}(V\ge495)$ 是多少？

对脱脂奶而言，495 是 $\mu - \frac{1}{2}\sigma$，所以 $\mathrm{P}(V \geq 495 \mid S) = b$（这正是第 (i) 部分的分母）。对全脂奶而言，495 是 μ，所以由对称性得 $\mathrm{P}(V \geq 495 \mid F) = \frac{1}{2}$。根据贝叶斯定理（Bayes' theorem），

$$\mathrm{P}(F \mid V \geq 495) = \frac{0.4 \times \frac{1}{2}}{0.6b + 0.4 \times \frac{1}{2}} = \frac{0.2}{0.6b + 0.2}.$$

令此式等于 $\frac{1}{3}$，得 $0.6b + 0.2 = 0.6$，所以 $b = \frac{2}{3}$。于是由 (1) 得 $4a = 7 - 4 = 3$，所以

$$a \approx \tfrac{3}{4}, \qquad b \approx \tfrac{2}{3}.$$

这些只是估计值，因为观察所得的比例只是真实概率的近似。

验算

- 估计值合情合理：$\frac{1}{2} < b < a < 1$，这是必然的，因为 $\mu < \mu + \frac{1}{2}\sigma < \mu + \sigma$。
- 代回检验：$0.6 \cdot \frac{2}{3} + 0.4 \cdot \frac{3}{4} = 0.4 + 0.3 = 0.7$，以及 $0.2/(0.4 + 0.2) = \frac{1}{3}$。
- 若分布为正态分布（Normal distribution），则 $a \approx 0.841$ 及 $b \approx 0.691$；以这些数值计算，第 (i) 部分得出 0.533 和 0.553。以（非正态的）拉普拉斯分布（Laplace distribution）作模拟，样本数目超过 10^5，结果与第 (i) 及 (ii)(a) 部分的所有公式吻合；例如 (ii)(a) 的模拟结果为 0.2518，而公式给出 0.2515。

常见错误

- 把 $\mathrm{P}(X \leq \mu - \frac{1}{2}\sigma) = b$ 当作正确，而其实应是 $1 - b$：对称性反射的是*尾部*，而不是累积概率。
- 忘记两种牛奶的平均值不同，所以同一个容量（例如 500 ml）对两种牛奶而言，与平均值相距的标准差（standard deviation）数目并不相同。
- 对两种牛奶都用 $\mathrm{P}(V \geq 495) = b$；对全脂奶而言，495 ml 正是平均值。

STEP 之外

这道题的要点在于：即使对分布一无所知，单凭对称性和全概率公式也可以走得很远；这里由两个总体混合而成的情况，是一个简单的*混合模型*（mixture model），而第 (ii)(b) 部分则是根据数据估计其参数的入门版本。另一个同样不依赖分布形状的相关结果是切比雪夫（Chebyshev）不等式，它只用方差（variance）便能给出 $\mathrm{P}(|X - \mu| \geq k\sigma)$ 的上界。可参阅 Blitzstein and Hwang, *Introduction to Probability*（概率论入门），或 Grimmett and Welsh, *Probability: An Introduction*（概率论导引）。

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。