详细解答

STEP II 2000，第 12 题

概率与统计 · 正态变量之和

> **STEP II 2000，第 12 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-q12

关键思路

Henry 迟到，正正就是他的总行程时间超过他在 0900 之前所余时间的时候，所以对每条路线，我们把两段独立的正态行程时间相加，得出单一个正态变量（normal variable）。标准化（standardising）之后，“迟到”就变成“一个标准正态变量超过某个数”，而由于 Φ 是递增的，两条路线毋须查表便可比较。余下的工作就是全概率定律（law of total probability）和贝叶斯定理（Bayes' theorem）。

解答

比较两条路线

我们假设在每条路线上，火车行程与其后的行程互相独立（independent）。

King's Cross。设 $X_1 \sim N(55, 25)$ 为火车行程时间，$Y_1 \sim N(30, 144)$ 为其后的行程时间。由于它们是独立的正态变量，它们的和也是正态的，而平均值（mean）与方差（variance）各自相加：

$$T_1 = X_1 + Y_1 \sim N(85,\ 169), \qquad \text{标准差13.}$$

火车在 0715 开出，所以 Henry 有 105 分钟，而当 $T_1 > 105$ 时他便会迟到。设 $Z \sim N(0, 1)$，则

$$\mathrm{P}(\text{迟到} \mid \text{King's Cross}) = \mathrm{P}\left(Z > \frac{105 - 85}{13}\right) = 1 - \Phi\left(\tfrac{20}{13}\right) = A.$$

Liverpool Street。同样地，$T_2 = X_2 + Y_2$，其中 $X_2 \sim N(65, 16)$ 与 $Y_2 \sim N(25, 9)$ 互相独立，所以

$$T_2 \sim N(90,\ 25), \qquad \text{标准差5.}$$

这班火车在 0720 开出，余下 100 分钟，所以

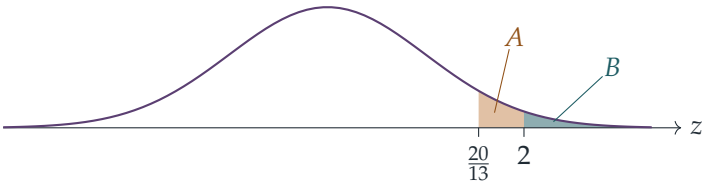
$$\mathrm{P}(\text{迟到} \mid \text{Liverpool Street}) = \mathrm{P}\left(Z > \frac{100-90}{5}\right) = 1 - \Phi(2) = B.$$

想一想 不查表，怎样比较 A 与 B?

Φ 是一个累积分布函数（cumulative distribution function），其密度为正，所以它是严格递增的。由于 $\frac{20}{13} < \frac{26}{13} = 2$，我们有 $\Phi(\frac{20}{13}) < \Phi(2)$，因此

$$A = 1 - \Phi\left(\tfrac{20}{13}\right) > 1 - \Phi(2) = B.$$

所以 Henry 若乘搭往 King's Cross 的火车，迟到的机会较大，这正是所要证明的。（有趣的是，这条路线平均而言反而较快；它之所以落败，是因为由 King's Cross 出发的那段地铁行程变化太大。）



标准正态曲线下涂上阴影的尾部：A 是整个阴影部分的面积，B 是 2 以外颜色较深的部分。

迟到次数的期望值

在单一次行程中，设 K 为 Henry 乘搭往 King's Cross 的火车这一事件，故 $\mathrm{P}(K) = p$，并设 D 为他迟到这一事件。根据全概率定律，

$$\mathrm{P}(D) = \mathrm{P}(D \mid K)\,\mathrm{P}(K) + \mathrm{P}(D \mid K')\,\mathrm{P}(K') = pA + (1-p)B.$$

假设 M 次行程互相独立，迟到次数便服从参数为 M 和 $\mathrm{P}(D)$ 的二项分布（binomial distribution），所以

$$L = M(pA + (1-p)B) = M(B + p(A-B)).$$

由于 $A - B > 0$ 且 $0 \le p \le 1$，表达式 $B + p(A-B)$ 随 p 递增，由 $p = 0$ 时的 B 增至 $p = 1$ 时的 A。因此对 p 的每一个可能值，

$$BM \le L \le AM,$$

左边的等号仅在 $p = 0$ 时成立，右边的等号仅在 $p = 1$ 时成立。用文字来说：L/M 是 A 与 B 的加权平均（weighted average），所以它必定介乎两者之间。

估计 p

在迟到的日子当中，他乘搭了往 King's Cross 火车的比例，是 $\mathrm{P}(K \mid D)$ 的估计值；而因为 M 很大，我们把观察所得的 $\frac{3}{5}$ 视为等于这个概率。根据贝叶斯定理，

$$\mathrm{P}(K \mid D) = \frac{\mathrm{P}(D \mid K)\,\mathrm{P}(K)}{\mathrm{P}(D)} = \frac{pA}{pA + (1-p)B}.$$

令此式等于 $\frac{3}{5}$：

$$5pA = 3pA + 3(1-p)B$$
$$2pA + 3pB = 3B,$$

所以估计值是

$$\hat{p} = \frac{3B}{2A + 3B}.$$

验算

- 以数值计算，$A \approx 0.0620$ 而 $B \approx 0.0228$，所以经 King's Cross 约有 6% 的时候迟到，经 Liverpool Street 则约 2%。对每条路线各模拟 4×10^5 次行程，结果吻合至小数点后三位。
- 代入这些数值，$\hat{p} \approx 0.355$。代回原式，$pA/(pA + (1-p)B) = 0.6$ 准确成立；而以这个 p 模拟整个过程，所得的迟到个案中约有 60% 来自 King's Cross。
- 对最后答案作合理性检查：假如两条路线同样可靠（$A = B$），迟到的情况便只反映他选乘哪班火车，所以当中有 $\frac{3}{5}$ 经 King's Cross，就表示 $p = \frac{3}{5}$。公式给出 $3B/5B = \frac{3}{5}$。由于 King's Cross 是较不可靠的路线，它在迟到的日子中所占的比例偏高，所以 p 必定小于 $\frac{3}{5}$，而事实正是如此。

常见错误

- 把标准差（standard deviation）相加，而不是把方差相加：King's Cross 路线的总时间方差为 $25 + 144 = 169$，所以其标准差是 13，而不是 $5 + 12 = 17$。
- 两班火车都用同一个 105 分钟的时限：往 Liverpool Street 的火车迟五分钟开出。
- 直接令 $p = \frac{3}{5}$。这个 $\frac{3}{5}$ 是 $\mathrm{P}(K \mid D)$，而不是 $\mathrm{P}(K)$；两者不同，是因为两条路线并非同样可靠。

STEP 之外

这里有两个想法，会在大学第一个概率课程中反复出现。独立正态变量之和仍是正态的这个事实，用矩母函数（moment generating function）来证明最为简洁；它也是正态分布族的稳定性（stability）的一个特例，而这种稳定性正是中心极限定理（central limit theorem）的基础。最后一部分则是统计推论（statistical inference）的缩影：我们观察到一个结果（迟到），然后运用贝叶斯定理推断隐藏的原因（选乘哪班火车）。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率入门）涵盖了这两个课题，而 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率导论）则有更多正是这种风格的贝叶斯题目。

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。