解答例

STEP II 2000、第 12 問

確率・統計 ・ 正規確率変数の和

> **STEP II 2000、問 12**
>
> 先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：
>
> step.maths.org/questions/00-s2-q12

鍵となる考え方

ヘンリーが遅刻するのは、移動時間の**合計**が 0900 までの持ち時間を超えるときであり、またそのときに限る。そこで各経路について、互いに独立（independent）な 2 つの正規分布（normal distribution）に従う所要時間を足し合わせ、1 つの正規確率変数にまとめる。標準化（standardising）によって「遅刻する」は「標準正規分布（standard normal distribution）に従う変数がある値を超える」と言い換えられ、Φ が増加関数であることから、数表を用いずに 2 つの経路を比較できる。残りは全確率の法則（law of total probability）とベイズの定理（Bayes' theorem）である。

解答

2 つの経路の比較

各経路において、列車での移動とその後の移動は独立であると仮定する。

キングス・クロス。列車の所要時間を $X_1 \sim N(55, 25)$、その後の所要時間を $Y_1 \sim N(30, 144)$ とする。これらは独立な正規確率変数なので、その和も正規分布に従い、平均と分散（variance）はそれぞれ足し合わされる：

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

列車は 0715 に発車するので、ヘンリーの持ち時間は 105 分であり、$T_1 > 105$ のとき遅刻する。$Z \sim N(0,1)$ とすると、

$$\mathrm{P}(\text{遅刻} \mid \text{キングス・クロス}) = \mathrm{P}\left(Z > \frac{105-85}{13}\right) = 1 - \Phi\left(\tfrac{20}{13}\right) = A.$$

リバプール・ストリート。同様に、$X_2 \sim N(65,16)$ と $Y_2 \sim N(25,9)$ は独立で $T_2 = X_2 + Y_2$ であるから、

$$T_2 \sim N(90,\,25), \qquad \text{標準偏差5.}$$

こちらの列車は 0720 に発車するので持ち時間は 100 分であり、

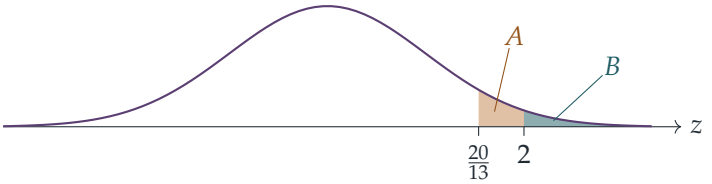
$$\mathrm{P}(\text{遅刻} \mid \text{リバプール・ストリート}) = \mathrm{P}\left(Z > \frac{100-90}{5}\right) = 1 - \Phi(2) = B.$$

考えてみよう 数表を使わずに A と B を比較するにはどうすればよいか？

Φ は正の確率密度をもつ累積分布関数（cumulative distribution function）なので、狭義単調増加である。$\frac{20}{13} < \frac{26}{13} = 2$ より $\Phi(\frac{20}{13}) < \Phi(2)$ となり、したがって

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

よって、題意のとおり、キングス・クロス発の列車に乗るほうがヘンリーは遅刻しやすい。（興味深いことに、この経路は平均的には**速い**。それでも不利になるのは、キングス・クロスからの地下鉄区間のばらつきが非常に大きいためである。）



標準正規分布の曲線の裾の部分に色を付けたもの：A は色の付いた部分全体の面積、B は 2 より右側の濃い部分の面積である。

遅刻回数の期待値

1 回の通勤について、ヘンリーがキングス・クロス発の列車に乗る事象を K とし、$\mathrm{P}(K) = p$ とする。また遅刻する事象を D とする。全確率の法則により、

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

M 回の通勤が互いに独立であると仮定すると、遅刻回数はパラメータ M と $\mathrm{P}(D)$ の二項分布（binomial distribution）に従うので、

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

$A - B > 0$ かつ $0 \le p \le 1$ より、式 $B + p(A-B)$ は p について増加し、$p = 0$ のとき B、$p = 1$ のとき A となる。したがって、p のとりうるすべての値に対して

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

が成り立ち、左側の等号は $p = 0$ のときに限り、右側の等号は $p = 1$ のときに限り成立する。言葉で言えば、L/M は A と B の加重平均（weighted average）なので、両者の間になければならない。

p の推定

遅刻したときのうちキングス・クロス発の列車に乗っていた割合は $\mathrm{P}(K \mid D)$ の推定値であり、M が大きいので、観測された $\frac{3}{5}$ をこの確率に等しいとみなす。ベイズの定理により、

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

これを $\frac{3}{5}$ に等しいとおくと：

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

よって推定値は

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

検算

- 数値的には $A \approx 0.0620$、$B \approx 0.0228$ であり、キングス・クロス経由では約 6%、リバプール・ストリート経由では約 2% の割合で遅刻する。各経路について 4×10^5 回の通勤をシミュレーションしたところ、小数第 3 位まで一致した。
- これらの値を用いると $\hat{p} \approx 0.355$ である。代入し直すと $pA/(pA + (1-p)B) = 0.6$ がちょうど成り立ち、この p で全過程をシミュレーションしたところ、遅刻のうち約 60% がキングス・クロス経由であった。
- 最後の答えの妥当性の確認：もし 2 つの経路の信頼性が等しければ（$A = B$）、遅刻の内訳は列車の選び方のみを反映するので、その $\frac{3}{5}$ がキングス・クロス経由であれば $p = \frac{3}{5}$ となるはずである。実際、公式は $3B/5B = \frac{3}{5}$ を与える。キングス・クロスは信頼性の低い経路なので、遅刻した日の中では実際より多く現れる。したがって p は $\frac{3}{5}$ より小さくなければならず、実際そうなっている。

よくある間違い

- 分散ではなく標準偏差（standard deviation）を足してしまう：キングス・クロス経由の合計時間の分散は $25 + 144 = 169$ なので、標準偏差は 13 であり、$5 + 12 = 17$ ではない。

- 両方の列車に同じ 105 分の持ち時間を使ってしまう：リバプール・ストリート発の列車は 5 分遅く発車する。
- いきなり $p = \frac{3}{5}$ としてしまう。$\frac{3}{5}$ は $\mathrm{P}(K \mid D)$ であって $\mathrm{P}(K)$ ではなく、2 つの経路の信頼性が等しくないため両者は異なる。

STEP のその先へ

ここで現れた 2 つの考え方は、大学初年次の確率論の講義を通じて繰り返し登場する。独立な正規確率変数の和が正規分布に従うという事実は、積率母関数（moment generating function）を用いると最もすっきりと証明でき、また中心極限定理（central limit theorem）の土台となる正規分布族の安定性の特別な場合でもある。最後の小問は統計的推測（statistical inference）の縮図である：ある結果（遅刻）を観測し、ベイズの定理を用いて隠れた原因（列車の選択）を推測したのである。Grimmett と Welsh による *Probability: An Introduction*（確率論入門）はこの両方を扱っており、Blitzstein と Hwang による *Introduction to Probability*（確率論入門）には、まさにこの種のベイズの問題がさらに数多く収録されている。

Beyond STEP 解答集：https://beyondstep.keithmcnulty.org/ja/。解答本文は CC0 1.0 である。問題そのものは © Cambridge University Press & Assessment に帰属し、ここには掲載していない。英語版から機械翻訳したものであり、まだ校閲を経ていない。