解答例

STEP III 2011、第 13 問

確率・統計 ・ 離散分布の最頻値

STEP III 2011、問 13

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/11-s3-q13

鍵となる考え方

離散確率変数（discrete random variable）の最も起こりやすい値を求めるとき、微分は使わない。隣り合う確率どうしを比べるのである。比（ratio）$\mathrm{P}(X=r+1)/\mathrm{P}(X=r)$ は r の簡単な有理関数であり、それがいつ 1 を超えるかを問うと、r についての 1 次不等式になる。比が 1 より大きい間は確率は増加し、それを下回ると減少に転じるので、最大値は切り替わる直前の r でとられる。復元抽出（with replacement）でも非復元抽出（without replacement）でも同じ方法が使え、変わるのは比だけである。

解答

小問 (i)

復元抽出では各回の取り出しは独立（independent）であり、それぞれ確率 $p=b/n$ で黒球となるので、$X\sim \mathrm{B}(k,p)$ である：

$$\mathrm{P}(X=r)=\binom{k}{r}p^r(1-p)^{k-r},\qquad r=0,1,\ldots,k.$$

まず $0<b<n$ と仮定する。このとき、これらの確率はすべて正である。$0\le r\le k-1$ に対して

$$\frac{\mathrm{P}(X=r+1)}{\mathrm{P}(X=r)}=\frac{k-r}{r+1}\cdot\frac{p}{1-p}=\frac{(k-r)\,b}{(r+1)(n-b)}.$$

分母は正であるから、この比が 1 を超えるのは次のときであり、かつそのときに限る：

$$\begin{aligned} & (k-r)b > (r+1)(n-b) \\ \iff \quad & kb - rb > rn - rb + n - b \\ \iff \quad & (k+1)b > (r+1)n \\ \iff \quad & r+1 < \frac{(k+1)b}{n}. \end{aligned}$$

同様に、

$$m = \frac{(k+1)b}{n}.$$

とおくと、比が 1 に等しいのは $r+1 = m$ のときに限り、1 より小さいのは $r+1 > m$ のときに限る。したがって、r が増加するにつれて、$\mathrm{P}(X = r)$ は到達先の値が $r+1 < m$ となる各段階では狭義に増加し、到達先が $r+1 > m$ となる各段階では狭義に減少する。$0 < b < n$ より $0 < m < k+1$ であるから、切り替わりは範囲 $0, \ldots, k$ の内部で起こることに注意する。

考えてみよう m が整数でないとき、最大の確率は正確にはどこでとられるか？

m が整数でないとき、$M = \lfloor m \rfloor$ とおくと、$M < m < M+1$ かつ $0 \le M \le k$ である。確率は $\mathrm{P}(X = M)$ まで狭義に増加し（各段階の到達先は $\le M < m$ である）、その後は狭義に減少する（以降の各段階の到達先は $\ge M+1 > m$ である）。よって、最も起こりやすい黒球の個数はただ一つで、

$$\left\lfloor \frac{(k+1)b}{n} \right\rfloor,$$

となり、示すべきことが示された。

一意でない場合。m が整数のとき、$r = m-1$ から $r = m$ への段階の比はちょうど 1 であるから $\mathrm{P}(X = m-1) = \mathrm{P}(X = m)$ となり、この等しい二つの値は他のどの値よりも大きい（確率は $m-1$ まで狭義に増加し、m の後は狭義に減少する）。$1 \le m \le k$ であるから、どちらもとりうる値である。したがって、答えが一意でないのは（$0 < b < n$ のもとで）$(k+1)b/n$ が整数であるときに限り、そのとき $m-1$ と m が同じく最も起こりやすい値となる。たとえば $n = 3$、$b = 1$、$k = 5$ のとき $m = 2$ であり、実際 $\mathrm{P}(X = 1) = \mathrm{P}(X = 2) = \frac{80}{243}$ である。

自明な場合はこの議論の外にある。$b = 0$ ならば確実に $X = 0$ であり、これは公式と一致する。$b = n$ ならば確実に $X = k$ であるが、上の比は $n-b$ で割ることを含むので、公式（これは $k+1$ を与えてしまう）は適用できない。

小問 (ii)

非復元抽出では、X は超幾何分布（hypergeometric distribution）に従う。標本を k 個の球からなる順序のない集合として選ぶと、$\binom{n}{k}$ 通りの集合はすべて同様に確からしいので、

$$\mathrm{P}(X=r)=\frac{\binom{b}{r}\binom{n-b}{k-r}}{\binom{n}{k}}.$$

$k\le b$ であるから、とりうる最大の値は $r=k$ である。黒でない球は $n-b$ 個より多くは取り出せないので、最小の値は $r_0=\max(0,k-(n-b))$ である。$r_0\le r\le k$ では確率は正であり、$r_0\le r\le k-1$ に対しては、$\binom{a}{r+1}=\binom{a}{r}\frac{a-r}{r+1}$ と $\binom{a}{s-1}=\binom{a}{s}\frac{s}{a-s+1}$ を用いて

$$\frac{\mathrm{P}(X=r+1)}{\mathrm{P}(X=r)}=\frac{b-r}{r+1}\cdot\frac{k-r}{n-b-k+r+1}.$$

この範囲ではどちらの分母も正であるから、比が 1 を超えるのは次のときであり、かつそのときに限る：

$$(b-r)(k-r)>(r+1)(n-b-k+r+1).$$

展開すると、左辺は $bk-r(b+k)+r^2$、右辺は $(r+1)(n-b-k+1)+r^2+r$ となる。r^2 の項は打ち消し合い、不等式は次のようになる：

$$\begin{aligned} & bk-r(b+k)>r(n-b-k+1)+(n-b-k+1)+r \\ \iff\quad & bk+b+k+1>r(n+2)+(n+2) \\ \iff\quad & r+1<\frac{(b+1)(k+1)}{n+2}. \end{aligned}$$

考えてみよう 切り替わる点は、とりうる値の範囲 $r_0,\dots,k$ の内部に本当にあるか？

$m=\frac{(b+1)(k+1)}{n+2}$ とおく。$b+1<n+2$ であるから $m<k+1$ である。下端については、$m>0$ は明らかであり、$r_0=k+b-n>0$ の場合には $(b+1)(k+1)>(n+2)(k+b-n)$ を示す必要がある。$x=n-b$、$y=n-k$ とおくと、左辺から右辺を引いたものは

$$(n+1-x)(n+1-y)-[(n+1)^2-1-(n+2)(x+y)]=1+x+y+xy>0.$$

よって、どの場合も $r_0<m<k+1$ であり、小問 (i) の議論がそのまま通用する。すなわち、確率は $r+1<m$ の間は狭義に増加し、$r+1>m$ となると狭義に減少する。したがって、一意である場合、最も起こりやすい黒球の個数は

$$\left\lfloor\frac{(b+1)(k+1)}{n+2}\right\rfloor.$$

である。一意でないのは $(b+1)(k+1)/(n+2)$ が整数のときに限る。そのとき $m-1$ と m（$r_0 \leq m-1$ かつ $m \leq k$ であるから、どちらもとりうる値である）は同じ確率をもち、他のどの値よりも起こりやすい。たとえば $n=10$、$b=7$、$k=5$ のとき $m=4$ であり、$X=3$ と $X=4$ が同率で、それぞれの確率は $\frac{105}{252} = \frac{5}{12}$ である。

検算

- コンピュータを用いて、すべての $n \leq 24$、$0 \leq b \leq n$、$1 \leq k \leq 19$（小問 (ii) では $k \leq b$）について両方の分布を分数で厳密に計算し、総当たりで最頻値（mode）を求め、すべての場合（小問 (ii) では 2565 通り）で両方の公式と両方の同率条件を確認した。
- $n=10$、$b=6$、$k=4$ のとき：$m = \frac{35}{12}$ なので公式は 2 を与え、実際 $X=2$ がただ一つの最頻値である。
- n が k に比べて非常に大きければ、数個の球を取り除いても黒球の割合はほとんど変わらないので、小問 (ii) は小問 (i) と同じようになるはずである。実際そうなっている：$(b+1)(k+1)/(n+2) \approx (k+1)b/n$。

よくある間違い

- 最頻値を平均の整数部分 $\lfloor kb/n \rfloor$ と結論してしまうこと。最頻値は平均に近いが、常に平均の整数部分に等しいとは限らない。
- 比がちょうど 1 のときには最頻値が一つではなく二つになることを忘れること。
- 小問 (ii) で $(r+1)(n-b-k+r+1)$ を展開する際に符号を誤ること。r^2 の項が打ち消し合うことは、よい確認になる。

STEP のその先へ

連続する確率どうしを比べることは、任意の離散分布の最頻値を求める標準的な方法であり、同じ計算により、小問 (i) の極限の場合としてポアソン分布（Poisson distribution）の最頻値 $\lfloor \lambda \rfloor$ が得られる。超幾何分布は標本理論（sampling theory）や、統計学におけるフィッシャー（Fisher）の正確検定（exact test）の中心にある。このように確率が増加してから減少する分布は単峰（unimodal）であるという。二項分布と超幾何分布の場合、この性質はより強い性質である対数凹性（log-concavity）から導かれる。さらに詳しくは、Grimmett and Welsh, *Probability: An Introduction*（確率論入門）や Blitzstein and Hwang, *Introduction to Probability*（確率論入門）を参照せよ。

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。