解答例

STEP III 2018、第 12 問

確率・統計 ・ 順序統計量

STEP III 2018、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/18-s3-q12

鍵となる考え方

事象「k 番目に小さい数が y 以下である」は、事象「n 個の数のうち少なくとも k 個が y 以下である」と同じであり、y 以下である値の個数は二項分布（binomial distribution）に従う。これにより、分布関数（distribution function）がただちに得られる。これを微分すると二つの和が現れ、(ii) の二つの二項係数の恒等式こそが、それらを望遠鏡和（telescoping sum）として打ち消し合わせる鍵となる。確率密度関数が分かれば、それを積分すると 1 になるという事実から、ある積分の値がただで求まる。さらに、同じ積分（n と k をずらしたもの）から平均が得られる。

解答

n 個の数を $X_1, \ldots, X_n$ とする。これらは互いに独立（independent）で、それぞれ $[0,1]$ 上の一様分布（uniform distribution）に従う。

小問 (i)

$0 \le y \le 1$ を満たす y を固定し、$X_i \le y$ となる X_i の個数を N とする。各 X_i は他と独立に確率 y で $X_i \le y$ を満たすから、$N \sim \mathrm{B}(n, y)$ である。

さて、$Y_k \le y$ となるのは、ちょうど少なくとも k 個の数が y 以下であるときである。実際、$Y_k \le y$ ならば、Y_k とそれより小さい $k-1$ 個の数はすべて y 以下であるから、$N \ge k$ である。逆に $N \ge k$ ならば、小さい方から k 個の数はすべて y 以下であり、特

に $Y_k \le y$ である。したがって

$$\mathrm{P}(Y_k \le y) = \mathrm{P}(N \ge k) = \sum_{m=k}^{n} \binom{n}{m} y^m (1-y)^{n-m},$$

となり、これが $(*)$ である。

小問 (ii)

階乗を書き下すと、$1 \le m \le n$ に対して

$$m\binom{n}{m} = \frac{n!}{(m-1)!\,(n-m)!} = n \cdot \frac{(n-1)!}{(m-1)!\,((n-1)-(m-1))!} = n\binom{n-1}{m-1},$$

であり、同様に $0 \le m \le n-1$ に対して

$$(n-m)\binom{n}{m} = \frac{n!}{m!\,(n-m-1)!} = n \cdot \frac{(n-1)!}{m!\,((n-1)-m)!} = n\binom{n-1}{m}.$$

である。（$m = n$ のとき左辺は 0 であり、$\binom{n-1}{n}$ を 0 と解釈すれば右辺も 0 である。）

確率密度関数。 Y_k の確率密度関数（probability density function）$f(y)$ は $(*)$ の導関数である。各項を積の微分法（product rule）で微分すると

$$f(y) = \sum_{m=k}^{n} \binom{n}{m} \left[m\, y^{m-1}(1-y)^{n-m} - (n-m)\, y^m (1-y)^{n-m-1} \right].$$

となる。後半の部分では、因数 $n-m$ があるため $m = n$ の項は消えるので、その和は実際には $m = n-1$ までである。二つの恒等式を用いると

$$f(y) = n \sum_{m=k}^{n} \binom{n-1}{m-1} y^{m-1}(1-y)^{n-m} \; - \; n \sum_{m=k}^{n-1} \binom{n-1}{m} y^m (1-y)^{n-1-m}.$$

? **考えてみよう** この二つの和はほとんど同じに見える。どうすれば両者をそろえられるだろうか？

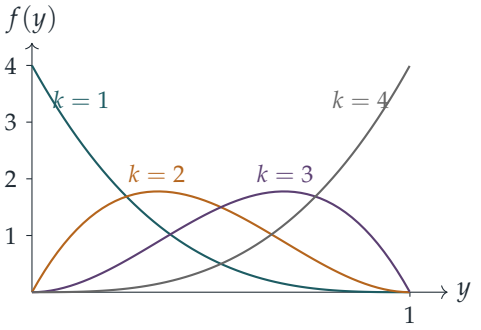
二つ目の和で $j = m+1$ とおくと、

$$n \sum_{j=k+1}^{n} \binom{n-1}{j-1} y^{j-1}(1-y)^{n-j},$$

となる。これは一つ目の和と項ごとに一致し、ただ k ではなく $k+1$ から始まる点だけが異なる。よって、一つ目の和の $m = k$ の項以外はすべて打ち消し合い、

$$f(y) = n\binom{n-1}{k-1} y^{k-1}(1-y)^{n-k},$$

を得る。これが示すべきことであった。（$k = n$ のときは、二つ目の和は空であり、一つ目の和は $m = n$ の項のみからなるので、この結果はやはり成り立つ。）



$n=4$ のときの確率密度関数。四つの数のうち k 番目に小さいものは $k/5$ の近くに集中する。

積分。 確率密度関数をその値域全体（ここでは $[0,1]$）で積分すると 1 になる。したがって

$$n\binom{n-1}{k-1}\int_0^1 y^{k-1}(1-y)^{n-k}\,\mathrm{d}y = 1,$$

であるから、

$$\int_0^1 y^{k-1}(1-y)^{n-k}\,\mathrm{d}y = \frac{1}{n\binom{n-1}{k-1}} = \frac{(k-1)!\,(n-k)!}{n!}.$$

である。これは $1 \le k \le n$ を満たすすべての整数について成り立つ。

小問 (iii)

定義より

$$\mathrm{E}(Y_k) = \int_0^1 y\,f(y)\,\mathrm{d}y = n\binom{n-1}{k-1}\int_0^1 y^{k}(1-y)^{n-k}\,\mathrm{d}y.$$

である。

考えてみよう この積分は、先ほど求めた積分とは少し違う。n と k をどのような値にすれば、その形になるだろうか？

被積分関数（integrand）は、$K=k+1$、$N=n+1$ として $y^{K-1}(1-y)^{N-K}$ であり、$1 \le K \le N$ であるから、(ii) の公式が適用できて

$$\int_0^1 y^{k}(1-y)^{n-k}\,\mathrm{d}y = \frac{k!\,(n-k)!}{(n+1)!}.$$

となる。$n\binom{n-1}{k-1} = \dfrac{n!}{(k-1)!\,(n-k)!}$ であるから、

$$\mathrm{E}(Y_k) = \frac{n!}{(k-1)!\,(n-k)!}\cdot\frac{k!\,(n-k)!}{(n+1)!} = \frac{k}{n+1}.$$

である。

検算

- $n = 1$、$k = 1$ のときは一様分布に従う数が一つだけであり、$\mathrm{E}(Y_1) = \frac{1}{2}$ となって、当然そうあるべき値になっている。
- 対称性（symmetry）：各 X_i を $1 - X_i$ に置き換えると、k 番目に小さい数は k 番目に大きい数になるから、$\mathrm{E}(Y_k) + \mathrm{E}(Y_{n+1-k}) = 1$ である。実際、$\frac{k}{n+1} + \frac{n+1-k}{n+1} = 1$ である。
- n 個の数は $[0,1]$ を $n+1$ 個の区間に分け、この答えは各区間の長さの平均が $\frac{1}{n+1}$ であることを示している。これは対称性から見て非常にもっともらしい。
- $n = 5$、$k = 2$ のとき：$y = 0.3$ での $(*)$ の値は 0.47178 であり、200 000 個の標本によるコンピュータシミュレーションでは 0.4726 が得られ、Y_2 の標本平均（sample mean）は $\frac{1}{3}$ に対して 0.33333 であった。

よくある間違い

- $\mathrm{P}(Y_k \le y) = \binom{n}{k} y^k (1-y)^{n-k}$ と書いてしまうこと。これは y 以下の値が**少なくとも** k 個ではなく、**ちょうど** k 個である確率である。
- 二つ目の和の $m = n$ の項が 0 であることを忘れ、望遠鏡和の打ち消し合いが一項ずれて成り立たないように見えてしまうこと。
- (iii) で積分を再利用するとき、k だけをずらして n をずらさないこと。

STEP のその先へ

Y_k は標本の**順序統計量**（order statistics）であり、Y_k はパラメータ k と $n-k+1$ のベータ分布（Beta distribution）に従う。(ii) の積分はベータ関数（Beta function）$B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}\,\mathrm{d}y = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ であり、ガンマ関数（Gamma function）とともに再び出会うことになるだろう。順序統計量は、標本中央値（sample median）やノンパラメトリック統計（non-parametric statistics）の基礎となっている。Blitzstein and Hwang, *Introduction to Probability*（確率論入門）は、順序統計量とベータ分布を見事に扱っており、ここでの議論とよく似た、ベータ積分の「ストーリー」による証明も載せている。

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。