2ch.net の投稿時系列データ解析と Pitman 分布の検証

弘前大理工 ^{A} Fintech Lab^{B}
守　真太郎 ^{A}, **久門　正人** ^{B}

Pitman sampling formula and an empirical study of choice behavior

^{A}*Hirosaki Univ.,* ^{B}*Fintech Lab.*
S. Mori^{A} **and M. Hisakado**^{B}

正の整数 r の分割に対する確率分布として 1 パラメータ ($\theta > 0$) の Ewens 分布 (Ewens sampling formula, ESF)、Ewens を 2 パラメータ ($0 \leq \alpha < 1, \theta > -\alpha$) に拡張した Pitman 分布 (Ewens-Pitman sampling formula,EPSF) が知られている。Pólya 壺の定式化では、初期状態として θ 個の黒玉を壺に用意し、壺からランダム（玉の個数に比例して）に取り出した玉が黒色なら新しい色の玉を黒玉とともに壺に、黒以外の球ならその玉の色を追加して壺に戻すとき、r 回の試行後の壺の中の r 個の新規に追加された玉の色の分布として Ewens 分布が導かれる。生物学の文脈では θ は突然変異による新種の追加に対応する。一方、α は黒（他の色の）玉を選択する確率を壺の中の玉の色の種類数 K_r に比例して増加（一定の割合で減少）させるパラメータであり、r 回試行後の r 個の追加された玉の色の分布が Pitman 分布に従う。我々はこの Pólya 壺モデルで壺からランダムに取り出される玉を直近の r 回の試行で追加された玉に限定するモデルでも Pitman 分布を定常分布として生成することを示した。この確率モデルを巨大掲示板 2ch.net でのスレッドの書き込みの時系列データで検証し、ニュース系掲示板での $r = 20 \sim 80$ の書き込み過程を記述することを示した。

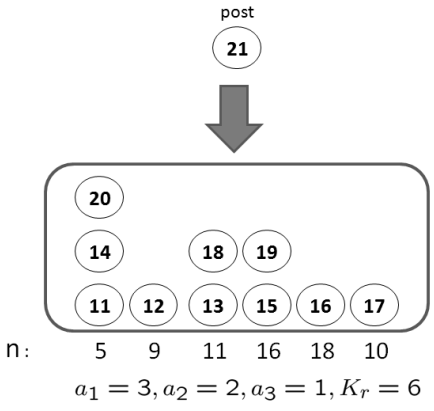

図 1　2ch.net の投稿過程と $r = 10$ の分割.21 番目の書き込みが過去 $r = 10$ 回の書き込みを参照して行われたとする。過去 10 回の書き込みは番号 5,9,11,16,18,10 の $K_r = 6$ 種のスレッドに投稿され、書き込み数 k のスレッド数 a_k は $a_1 = 3, a_2 = 2, a_3 = 1$.

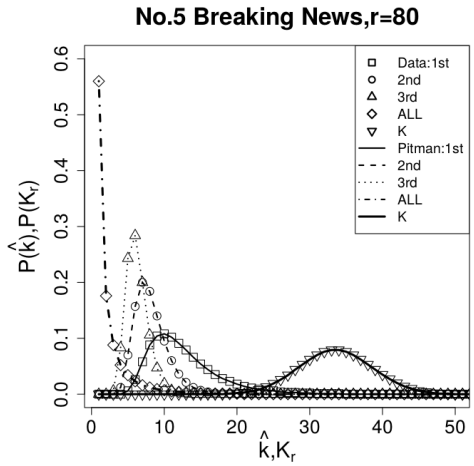

図 2　$r = 80$ としたときの K_r 種のスレッドの書き込み数 $\hat{k}_j, j = 1, \cdots, K_r$ を降順に並べたときの $k_1(\square) \geq k_2(\circ) \geq k_3(\triangle)$ と $k_j(\Diamond)$ および $K_r(\bigtriangledown)$ の分布と Pitman 分布との比較.