平等度の統計学と確率的順位付け

服部哲弥　（慶應・経済）

2012.08.10 集中講義「統計学」
北海道大学理学部数学科

START

０－１．背景－べき法則

注：前半は「統計学」講義の続き．後半が私の研究

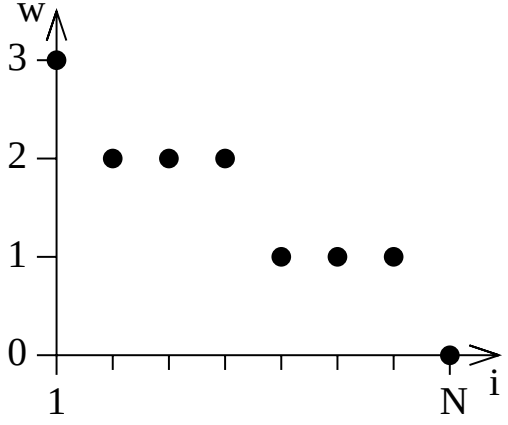
講義の例：公平な硬貨投げ３回中おもての枚数の分布 $B(3, 0.5)$

大きさ $N = 8$ のサンプルデータがこの分布どおり出現したとする

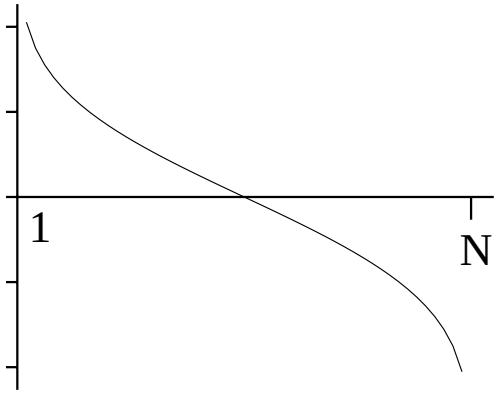
おもて	３	２	１	０
件数	１	３	３	１

見方をかえて，データ間のおもての枚数の異同に注目

データをおもて枚数の大きい順に並べて $w_i,\ i = 1, 2, \ldots, N$



枚数対順位



ガウス分布の分布関数の逆関数

指数法則

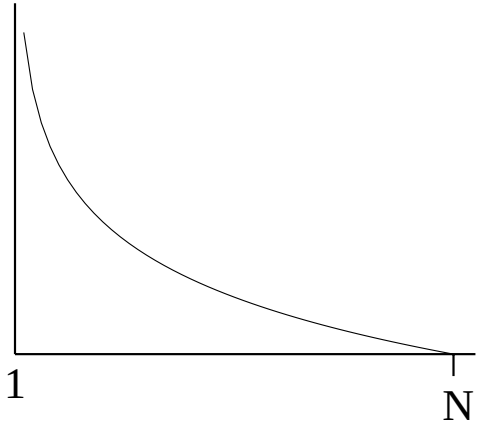
順序・順位：データや母分布を講義と異なる視点→おもしろい

- **順位iの値w_i, $i = 1, 2, \ldots, N$（非増加列）**
- **社会的な分布（例：世帯所得）：正$w_i > 0$**
- **簡単（密度と順位とも数式が簡単な分布は少ない）**

例. $w_i = c \log \dfrac{N}{i}$, $i = 1, 2, \ldots, N$, $c > 0$

分布関数：$\mathrm{P}[\,[x, \infty)\,] = \frac{1}{N}\sharp\{i \mid w_i \geqq x\} = \frac{1}{N}\sharp\{i \mid i \leqq Ne^{-x/c}\} \sim e^{-x/c}$

指数分布（「幾何分布」）



格差の度合いをパラメータとしたい→講義になかった例を探す

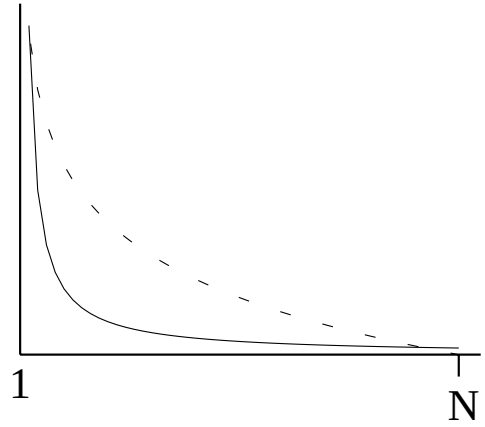
（注：後半でX_iを別の意味に使うのでw_iとしています．xとの混在すみません）

べき法則

べき法則（一般化Zipfの法則，一般化Paretoの分布，対数線形分布，…）

$$w_i = a\left(\frac{N}{i}\right)^{1/b},\ i = 1, 2, \ldots, N,\ a, b > 0$$

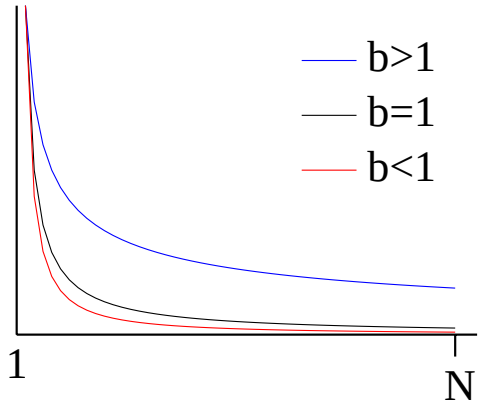
$$\mathrm{P}[\,[x, \infty)\,] = \frac{1}{N}\sharp\{i \mid w_i \geqq x\} = \left(\frac{a}{x}\right)^b$$



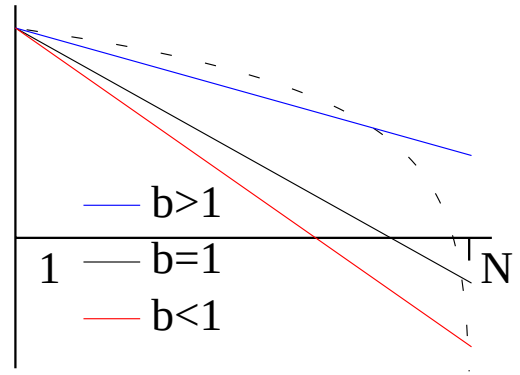
平等度のパラメータ

$$w_i = a\left(\frac{N}{i}\right)^{1/b} \Leftrightarrow \log w_i = \frac{1}{b}\log\frac{N}{i} + \log a$$

a: **最低所得，** b: **平等度** $\frac{w_1}{w_N} = N^{1/b}$ **：b大のとき平等**



w_i vs i



$\log w_i$ vs $\log i$

下位（ロングテール＝長い尾）が新しい研究対象なので，事実認識には左図が良い（両対数は指数を見極める上で有効）

０－２．流行度の順位付け

べき分布の社会・経済現象への当てはめ

古典：**所得格差**（Pareto）

大きな注目：大きな事故や損害が正規分布より頻繁（株価下落）**「黒鳥」**

現代から近未来の注目：**ロングテールビジネス**の可能性

・多様な商品や意見への購入者・支持者の分布＝**流行度**の分布

・**ウェブ**（2.0）時代→カテゴリごとの選択肢の急増，コストの急減

→現代から近未来に重要度が飛躍的に増加する可能性

ウェブ時代の巨大な選択肢と流行度

- 巨大掲示板群2ch.netの特定の板のスレッド一覧（関心テーマ）
- ブログ集合体amebloのブログ人気ランキング（発信者＝αブロガ）
- 動画投稿サイト**ニコニコ動画**，You Tubeの再生数（映像娯楽）
- ダウンロードサイトiTunesのDL数（音楽娯楽）
- アイドル集合体AKB48の人気投票（アイドルの「消費」）
- オンライン書店Amazonの書籍売上ランキング（知識／娯楽）
- 各種通販サイト**楽天**等の商品種別ごとの商品売上（商品）
- 論文**リポジトリ**のDL数，論文の引用数ランキング（研究成果）

インターネット時代に初めて可能になった項目の多さ

人気は集中し，目立たない多数の項目（しかし売上が正の「**長い尾**」）

参考：自然現象におけるべき分布の例に強度上の予震回数の分布

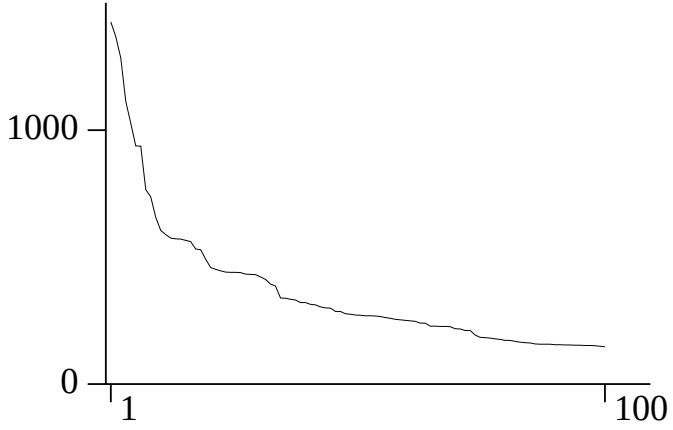
ニコニコ動画

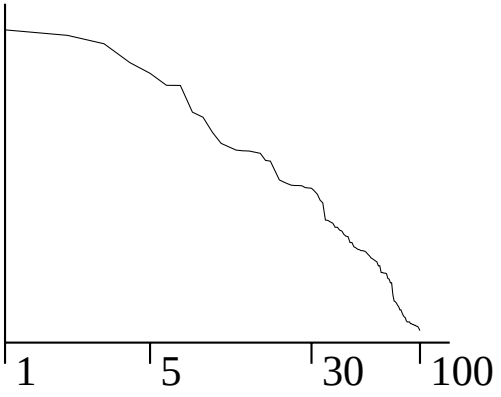
困難：長い尾の実現したweb時代だが，下位の値w_iは殆ど公開されない

下位の順位詳細には興味は集まらない→コストに比べ宣伝効果がない

2011年度慶應・経済3年山下紘史研究プロジェクト（ニコニコ動画，AKB48選挙）

ニコニコ動画再生回数（総合，1時間）



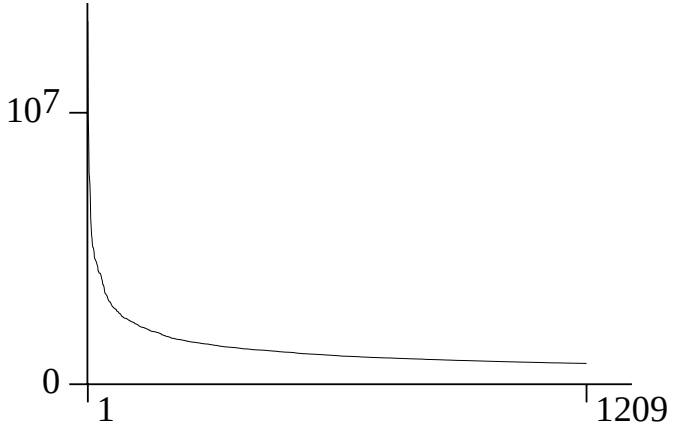


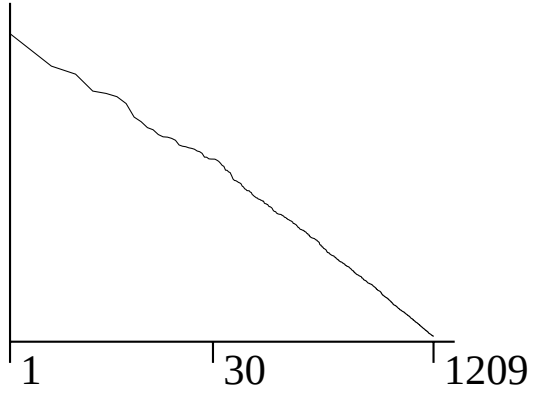
$b = 2.3$ $(i \leqq 29)$ $b = 1.1$ $(29 < i \leqq 100)$

指数分布よりはべき分布（特に「尾」の部分）

ニコニコ動画（続き）

ニコニコ動画再生回数（総合，累積）

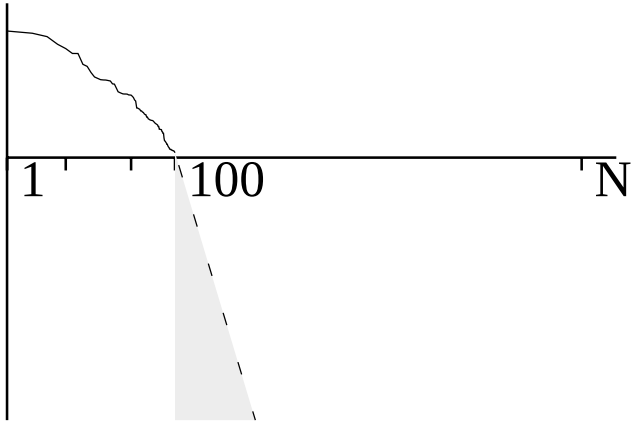




$b = 2.8\ (i \leqq 28)$ $b = 2.2\ (28 < i \leqq 1209)$

長い尾のデータの推定

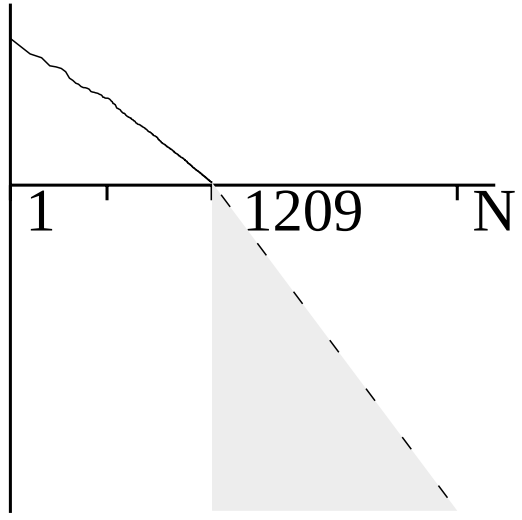
中・下位の個別の動画再生数（ランキング）は非公表　→　べき法則を仮定すれば，動画総再生数と動画総数から下位の分布も決まる



$N = 7 \times 10^6$, $\sum_i w_i = 4.8 \times 10^4$,

$b = 0.44$

($w_1 = 3697$, $w_{100} = 147$, $\sum_{i>100} 1.2 \times 10^4$)



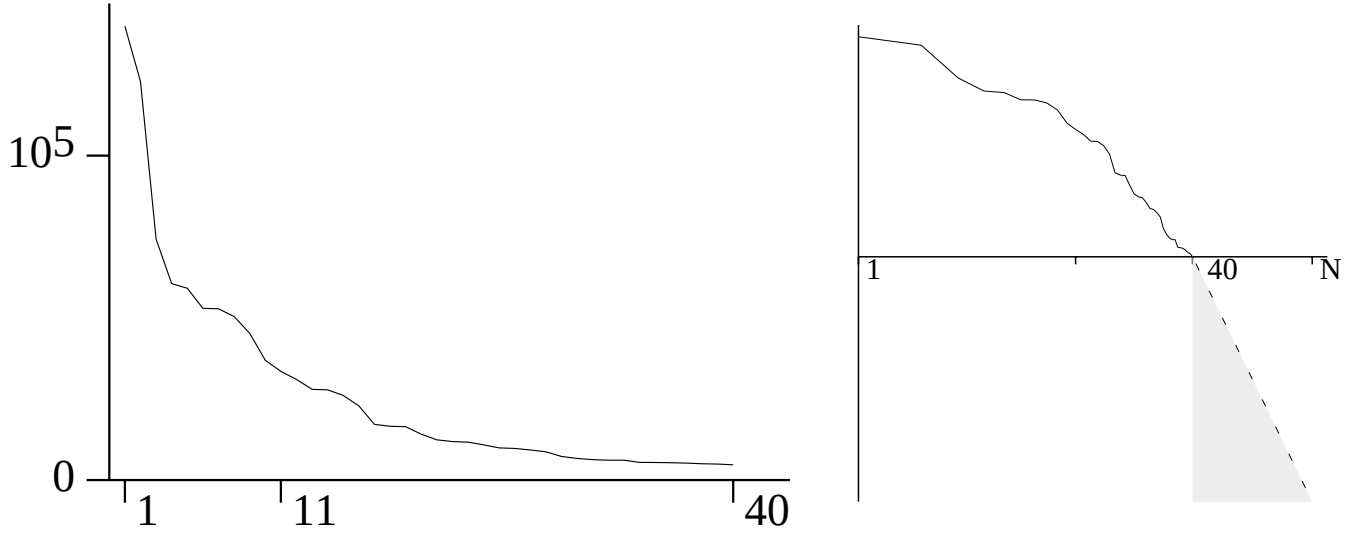
$N = 7 \times 10^6$, $\sum_i w_i = 2.8 \times 10^{10}$,

$b = 1.3$

($w_1 = 1 \times 10^7$, $w_{1209} = 8 \times 10^5$, $\sum_{i>1209} 2.8 \times 10^{10}$)

2011年AKB48総選挙（人気投票）

$N = 150$, **総投票数**120**万**，$b = 1.7$ $(i \leqq 11)$,
$b = 0.59$ $(11 < i \leqq 40)$, $b = 0.35$ $(40 < i \leqq 150)$



ニコ動もAKB選挙も，ヘッド（上位）と累積はb大$(b > 1)$だが，テール（長い尾）で短期（流行度）は$b < 1$

指数法則よりはべき法則だが，その中では（格差＝テールの打ち切り）

ロングテールビジネス分析の難しさ

下位データは公開されないため，ロングテール部分のデータ入手は**困難**

正しい応用数学：**現実における困難**の所在が分かれば，そこに**新しい数学**を投入できる（べき法則の仮定下で，総数が分かれば下位も決まる）

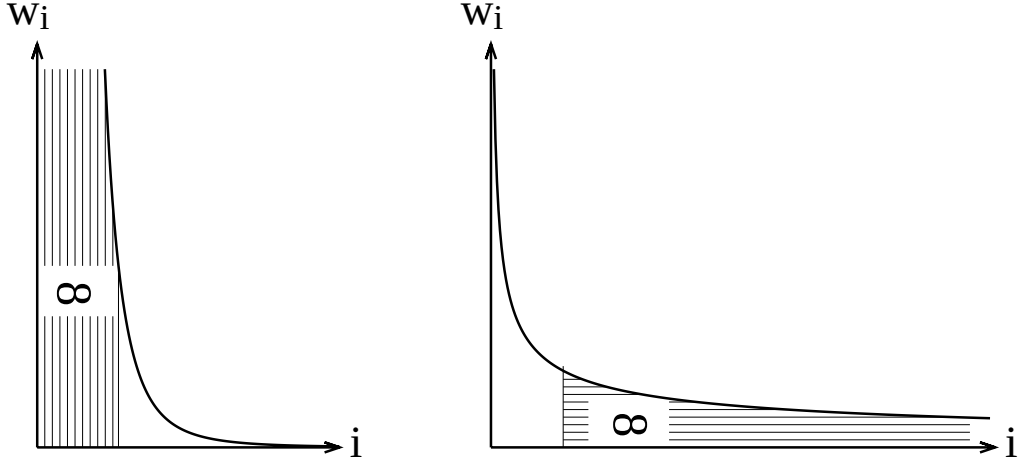
商売（アマゾン書店）：売上は**秘匿**するのが普通→総数もわからない
ロングテール部分の推測方法？

★**後述**：ランキングから分布を逆算できる！（**確率順位付け模型**）

以下，**簡単のため**べき法則を仮定して，その意味をまず確認する

０－３．指数b

多品目商品を扱う小売業．例：大型書店．商品１点⇔書籍１タイトル
一定期間の売上順に商品を並べて，i番目の商品の売上をw_iとおく



横軸 ⇔ i ⇔ 商品１点

縦軸 ⇔ w_i ⇔ 商品iの単位時間当売上

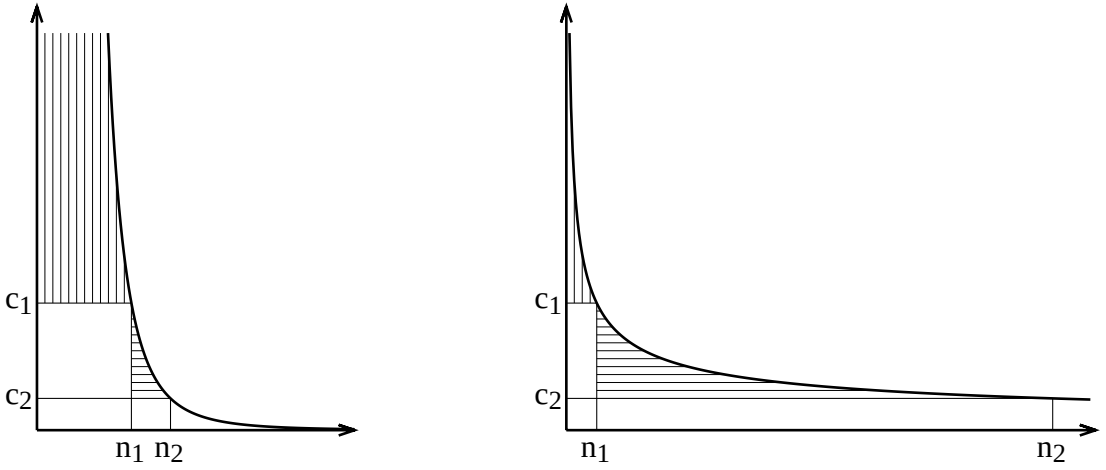
左図：$b < 1$格差大　右図：$b > 1$平等に近い

$b > 1$のとき，「尾」（非売れ筋）の売上合計が無視できない
（ロングテールビジネスモデル成立の可能性）

べき法則とロングテールビジネス

i ⇔ 横軸 ⇔ 商品1点（本のタイトル），w_i ⇔ 縦軸 ⇔ 商品iの単位時間当売上

インターネットリテール：インターネットの日常化で**現実的**(Web2.0)
商品「陳列」コスト激減（地代，盗難破損劣化，棚卸し）
ロングテールビジネスが初めて可能に　→　成立・規模の分析の必要



$b > 1$（右図）のとき，低コスト$(c_2 \ll c_1)$ならば，**ロングテール**ビジネスモデル成立の可能性

ロングテールビジネスの可能性の数理

- $b < 1$ → 商品点数が劇的に増えても，売上への貢献は微小
- $b > 1$ → 商品点数が劇的に増えれば，市場は拡大する期待

◯近未来（ネット小売り始まりの時代）
→ $b \sim 1$ならば，$b > 1$と同様，種々の対象を開拓する余地
べき法則の仮定の補正：コスト，w_iの$(\log i)^2$の項の係数が負（予想）
◯中長期→$b \sim 1$ならば，$b < 1$と同様，飽和は早い

インターネットリテールの将来予測に役立つ

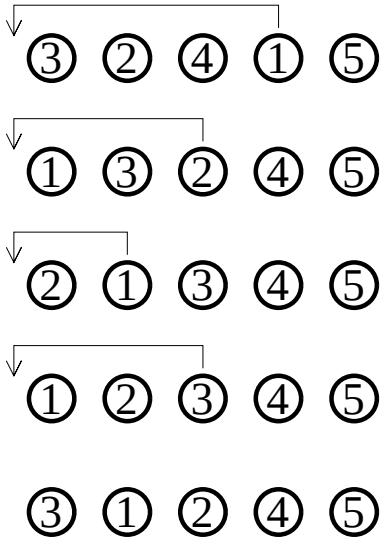
困難（再掲）：テールは関心が薄い → データが見あたらない
商売（アマゾン書店）では特に，売上は秘匿するのが普通
★ロングテール部分の推測方法？鍵：アマゾン書店はランキングを公開

１－１．先頭に跳ぶ規則

ここから研究の本題．ここまでの話を一度忘れて，数学的な模型
（のちほど，一見無関係な０．と１．が結びつくことを言います）

N **自然数**

N **個の粒子を一列に並べた系の並び方についてのランダムな時間発展**



どれかの粒子がランダムに先頭に跳び，追い越された粒子は順位を 1 **ずつ下げる**

例：粒子 1, 2, 1, 3 **の順に跳んだときの並び方の時間発展**

先頭に跳ぶ規則 M.L.Tsetlin (1963)**，**

積ん読，超整理法，最後に跳んだ順，…

確率順位付け模型

個々の粒子の順位（ランキング）の時間発展（確率過程）に注目
ゼッケン番号iの粒子の時刻tでの位置（ランキング）$X_i^{(N)}(t)$

$t \geqq 0,\ i = 1, 2, \ldots, N,\ X_i^{(N)} :\ \Omega \times \mathbb{R}_+ \to \{1, 2, \ldots, N\}$

$X^{(N)} = (X_1^{(N)}, \cdots, X_N^{(N)})$

初期値： $X_i^{(N)}(0) = x_i^{(N)},\ i = 1, 2, \cdots, N$

時間発展：Poisson**のランダムな測度に応じた時刻で** $X^{(N)}(t)$ **は変化**

(0) **先頭に跳ぶ規則：ある粒子が先頭**$(X_i^{(N)}(t) = 1)$**に跳び，追い越**

された粒子は順位を1**ずつ下げる**$(X_j^{(N)}(t) = X_j^{(N)}(t-) + 1)$

→先頭に跳ぶランダムな時刻で決まる：できる限り簡単なものを選ぶ

時間発展

(1) 粒子iが時間$(s,t]$の間に先頭に跳ぶ回数$\nu_i^{(N)}((s,t])$は
平均$\lambda = w_i^{(N)}(t-s)$のポワッソン分布$\mathrm{P}[\,\nu_i^{(N)}((s,t]) = k\,] = e^{-\lambda}\dfrac{\lambda^k}{k!}$

(2) 下位の粒子が1位に跳ぶ毎に1ずつ順位が下がる

- $X_i^{(N)}(t)$＝最後に跳んだ順に並べるときの位置（順位）

(3) 先頭への跳びは粒子間で独立，異なる時間で独立
（ポワッソン過程．$\nu_i^{(N)}((1,2]) \perp \nu_i^{(N)}((3,4])$）

- 確率1で同一時刻に先頭に跳ぶのは1個

確率微分方程式（参考）

N 固定．$i = 1, 2, \ldots, N$

$\nu_i^{(N)} : (\Omega, {\mathbb{R}_+}^2) \to \mathbb{Z}_+$ は i について独立な，

$ds\, d\xi$ を強度とする増分 1 のポワッソン過程

$$X_i^{(N)}(t) = x_i^{(N)}$$
$$+ \sum_{j=1}^{N} \int_{s\in(0,t]} \int_{\xi\in\mathbb{R}_+} 1_{X_i^{(N)}(s-)<X_j^{(N)}(s-)}\ 1_{\xi\leqq w_j^{(N)}(X_j^{(N)}(s-),s)}\ \nu_j^{(N)}(d\xi ds)$$

（↑下位の粒子の先頭へのジャンプに押される変化）

$$+ \int_{s\in(0,t]} \int_{\xi\in\mathbb{R}_+} (1 - X_i^{(N)}(s-))\ 1_{\xi\leqq w_i^{(N)}(X_i^{(N)}(s-),s)}\ \nu_i^{(N)}(d\xi ds)$$

（↑先頭へのランダムなジャンプ）

1_A は事象 A（または集合 A）の定義関数

- $w_i^{(N)}$ たちは，時刻依存性（昼夜差）を持つ場合 [針谷服部服部永幡竹島小林]，位置依存性（先頭注目効果）を持つ場合 [服部楠岡]，に拡張できる

１－２．　大数の法則

$J_i^{(N)}(0,t) = \{\nu_i^{(N)}((0,t]) > 0\}$　　（粒子iが時刻tまでに先頭に跳ぶ事象）

特性曲線（の，離散版）　$Y_C^{(N)}(t) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} 1_{J_i^{(N)}(0,t)}$

先頭に跳ぶ以外で追越し無し→$Y_C^{(N)}$はジャンプ済み粒子と未ジャンプ粒子の境界

命題． $\lambda^{(N)} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \delta_{w_i^{(N)}} \to \exists\lambda\ (N \to \infty)$ **ならば，**

$Y_C^{(N)}(t) \to y_C(t) = 1 - \displaystyle\int_{\mathbb{R}_+} e^{-w\,t}\lambda(dw)\ (N \to \infty$，**概収束**$)$　◇

証明：　独立確率変数の和の大数の強法則：$Y_C^{(N)}(t) - \mathrm{E}[\,Y_C^{(N)}(t)\,] \to 0$

＋ポワッソン分布：$\mathrm{P}[\,J_i^{(N)}(0,t)\,] = 1 - \mathrm{P}[\,\nu_i^{(N)}((0,t]) = 0\,] = 1 - e^{-w\,t}$　□

位置ジャンプ率結合経験分布の収束

ジャンプ率$w_i^{(N)}$と位置$Y_i^{(N)} = \frac{1}{N}(X_i^{(N)} - 1)$の結合経験分布：

$$\mu_t^{(N)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{(w_i^{(N)}, Y_i^{(N)}(t))}$$

定理. $\mu_0^{(N)} \to \exists \mu_0$ $(N \to \infty)$ ならば，各$t > 0$に対して$\mu_t^{(N)} \to \mu_t$ $(N \to \infty$，概収束）.

μ_tは（初期分布μ_0を用いてあらわに書ける）非ランダムな分布. ◇

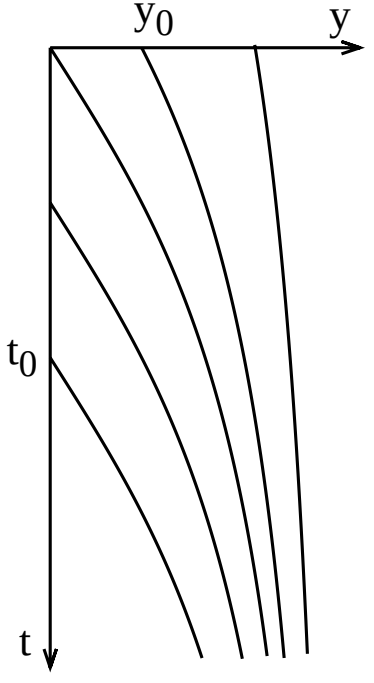
skip

証明の鍵　－　特性曲線$y_C(y_0, t_0; t)$

$\Gamma_i = \{(y, 0) \in [0, 1) \times \mathbb{R}_+ \mid y \geqq 0\}$（初期点集合），
$\Gamma_b = \{(0, t) \in [0, 1) \times \mathbb{R}_+ \mid t \geqq 0\}$（境界点集合），$\Gamma = \Gamma_i \cup \Gamma_b$
$(y_0, t_0) \in \Gamma$ に $Y_C^{(N)}(t) = Y_C^{(N)}(y_0, t_0; t)$ と $y_C(t) = y_C(y_0, t_0; t)$ を拡張

$$Y_C^{(N)}(y_0, t_0; t) := y_0 + \frac{1}{N} \sum_{i;\, Y_i^{(N)}(t_0) \geqq y_0} 1_{J_i^{(N)}(t_0, t)}$$

$$y_C(y_0, t_0; t) := 1 - \int_{\mathbb{R}_+} e^{-w\,(t-t_0)} \mu_{t_0}(dw \times [y_0, 1))$$



大数の法則（独立確率変数列）：
$Y_C^{(N)}(y_0, t_0; t) \to y_C(y_0, t_0; t)$

証明の鍵　ー　特性曲線に沿って分布を見る

$$U^{(N)}(dw, y, t) = \mu_t^{(N)}(dw \times [y, 1)) = \frac{1}{N} \sum_{i;\ Y_i^{(N)}(t) \geqq y} \delta_{w_i^{(N)}}(dw)$$

（再掲） $Y_C^{(N)}(y_0, t_0; t) = y_0 + \frac{1}{N} \sum_{i;\, Y_i^{(N)}(t_0) \geqq y_0} 1_{J_i^{(N)}(t_0, t)}, \quad (y_0, t_0) = \gamma \in \Gamma$

補題． $\varphi^{(N)}(dw, \gamma, t) = U^{(N)}(dw, y_C^{(N)}(\gamma, t), t)$ **は独立確率変数の大数の法則によって** $N \to \infty$ **で収束** ◇

証明：　先頭に跳ばない限り粒子間は追い越し無し　→　** $y_C^{(N)}$ **を通しての粒子の出入り無し　→　** $y_C^{(N)}$ **より右の粒子が先頭に跳ぶ効果のみで時間発展 □

$y = y_C(\gamma, t) = y_C(y_0, t_0, t)$ **の** γ **についての逆関数** $\widehat{\gamma}$ **をとれば，** $U(dw, y, t) = U(dw, y_C(\widehat{\gamma}(y, t), t) = \varphi(dw, \widehat{\gamma}(y, t), t)$ **への収束を得て，定理の証明を終える** □

μ_t

$t > 0$ fix

$y = y_C(\gamma, t)$, $\gamma \in \Gamma_t = \Gamma_i \cup \{(0, t_0) \in \Gamma_b \mid t_0 \leqq t\}$ **の逆関数**

$\widehat{\gamma}(y, t) = (y_0(y, t), t_0(y, t)) = \gamma$

$\mu_t(dw \times [y, 1))$

$= e^{-w\,(t - t_0(y,t))}\, \mu_0(dw \times [y_0(y, t), 1))$

$$= \begin{cases} e^{-w\,t}\, \mu_0(dw \times [y_0, (y, t), 1)) & y > y_C(\gamma, t) \\ e^{-w\,(t - t_0(y,t))}\, \mu_0(dw \times [0, 1)) & y < y_C(\gamma, t) \end{cases}$$

$(\lambda(dw) = \mu_0(dw \times [0, 1)))$

極限を記述する偏微分方程式

ジャンプ率が有限種類： $\lambda = \sum_{\beta} r_\beta \delta_{w_\beta},\ \sum_{\beta} r_\beta = 1,\ w_\beta, r_\beta > 0$

$u_\alpha = \mu_0(\{w_\alpha\} \times \cdot) :\ [0,1) \to \mathbb{R}_+$: **非負滑らか狭義減少，のとき**

定理． $U_\alpha(y,t) = U(\{w_\alpha\}, y, t)$ **は次の偏微分方程式系の初期値問題の時間大局的一意古典解：**

$$\frac{\partial U_\alpha}{\partial t}(y,t) + \sum_{\beta} w_\beta U_\beta(y,t) \frac{\partial U_\alpha}{\partial y}(y,t) = -w_\alpha U_\alpha(y,t)$$

境界条件 $U_\alpha(0,t) = r_\alpha$**，初期値** $U_\alpha(\cdot,0) = u_\alpha(\cdot),$ ◇

2－1．Amazonランキングの謎

Amazon.co.jp

本のページ中程やや下
Amazon.co.jp **ランキング**

「Amazonの謎順位。」

‘Internet retailers are extremely hesitant about releasing specific sales data’

謎ならば答えを探そう

ランキングの時間変化

・本を書くと，自分の本の順位が気になる．

Amazon.co.jp ランキング：本で373,406位

(1時間後) Amazon.co.jp ランキング：本で373,977位

(2 時間後) Amazon.co.jp ランキング：本で374,693位

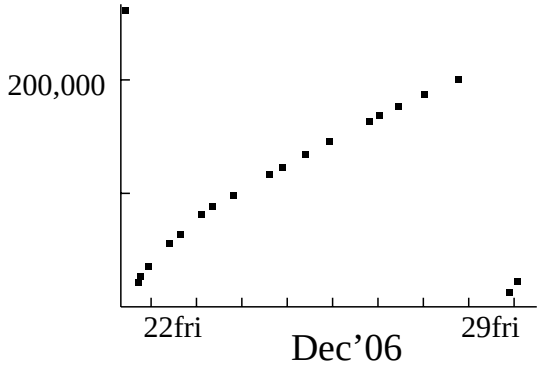
・が，漫然と眺めても１時間に１度更新されることくらいしか分からない

・方針を逆転させ，先に簡単な数理モデルを立てて，データの特徴を捉えているか付き合わせる．

ランキング＝流行度を反映する順位→理想化：

売れた瞬間を１位にする先頭に跳ぶ規則（確率順位付け模型）

アマゾン書店での順位の大きな跳び



１回の更新での大きな順位改善の確認

- 先頭に跳ぶ規則は少なくとも定性的にアマゾン書店ランキングを説明

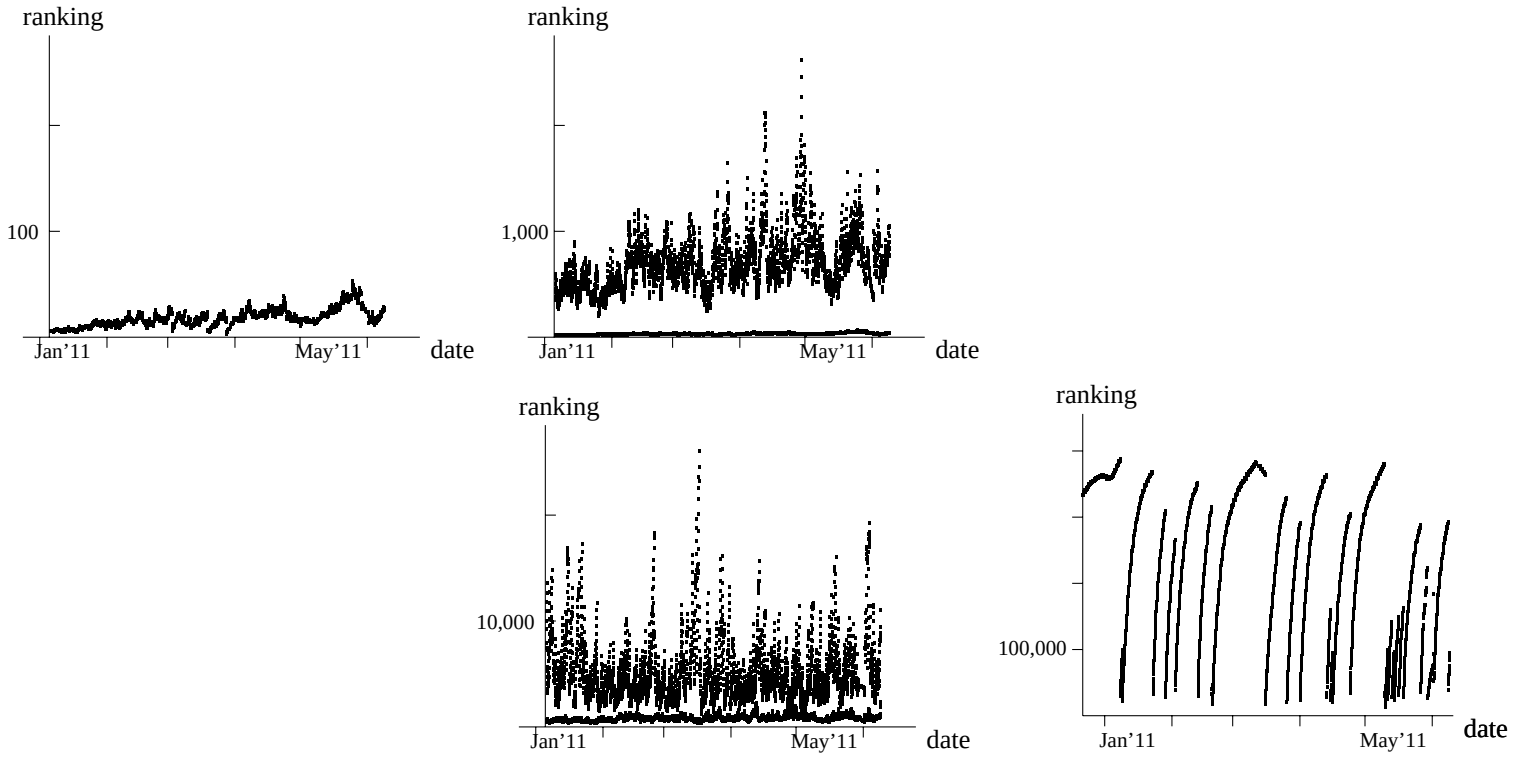
あり得る疑問：

- 順位はもっと安定的（確定的）な数値であるべきだろう
- アマゾン書店がそんな単純なアルゴリズムですますまい

→定量的には到底役立たない？　⇔　★事実：合う（後述）

- 順位が安定しているのはよく売れる一部の上位だけ．大多数の本は数ヶ月に１回以下しか売れない．それらの本の流行度の判断基準は直近の１回の購入しかない．
- アマゾン書店のアルゴリズムは，単純な先頭に跳ぶ規則でなく，少なくとも数時間前までの購入状況も反映．しかし，数ヶ月に１回以下しか売れない大多数の本にとっては関係ない．

大数の法則とサンプル



ビッグヒット vs. 普通の本（ロングテール側）
巨大な系の流行度の順位は，大多数にとって確率的（普遍性）

２－２．大数の法則とランキング

前掲命題． $x_i^{(N)} = X_i^{(N)}(0) = 1$ **のとき，次に先頭に跳ぶまで**

$$X_i^{(N)}(t) = X_C^{(N)}(t) + 1 \sim N\, Y_C^{(N)}(t) \sim N - N \int_{\mathbb{R}_+} e^{-wt} \lambda(dw)$$

w_i **⇔商品** i **の平均注文頻度，** λ **⇔注文頻度の分布**

限られたデータによって，社会についての何を知るべきか？

λ **が（一般化）** Zipf **の法則** (Pareto **分布**) **の場合**

$$w_i^{(N)} = a \left(\frac{N}{i}\right)^{1/b} ;\ a\text{:}\ \textbf{最低収入，}\ b\text{:}\ \textbf{平等性の指数}$$

$$X_i^{(N)}(t) \sim N - Nb(at)^b \Gamma(-b, at));\ \Gamma(z,p) = \int_p^\infty e^{-x} x^{z-1}\, dx$$

N, a, b **を与えれば決まる（データを使って統計的当てはめ）**

注． N **は** Amazon **公称ではなく実効値** (Pareto **分布に近い範囲**)

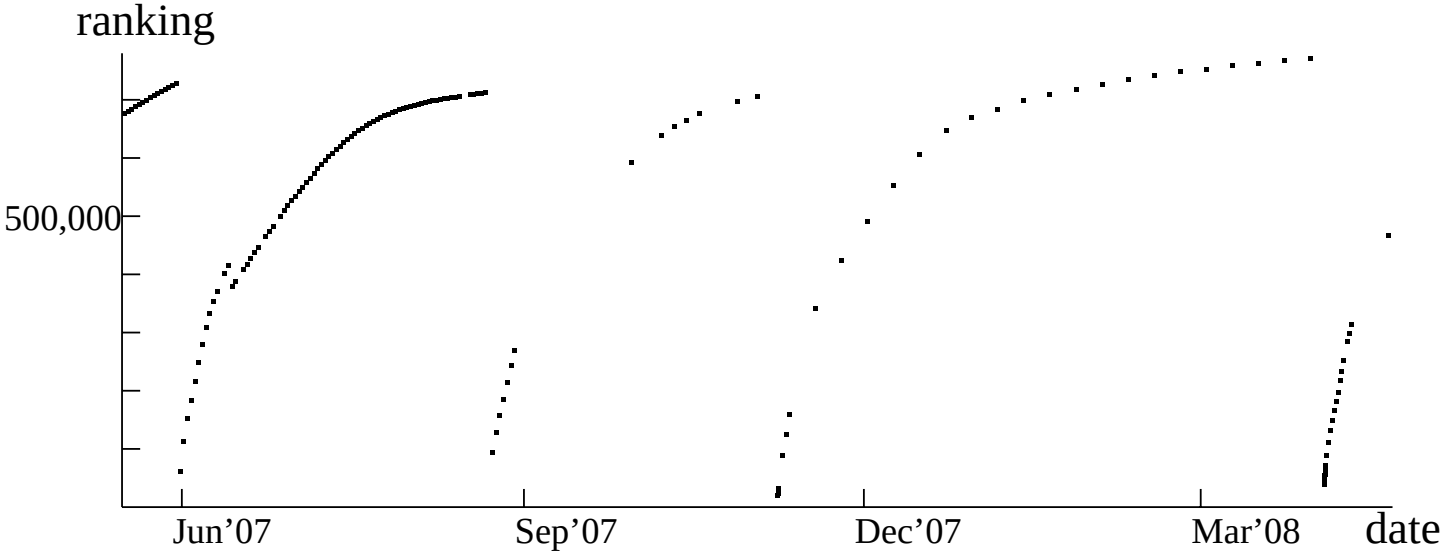
最小2乗法

データ：(t_ℓ, x_ℓ), $\ell = 1, 2, \ldots, n_d$
いいかどうかは別にして，理論とのずれは独立なポワッソンと仮定すると

$$E = E(N, a, b) = \sum_{\ell=1}^{n_d} (x_\ell - N\, y_C(t_\ell))^2 / x_\ell \sim \sum_{\ell=1}^{n_d} N[0,1]^2$$

- **注文のばらつきは大数の法則で処理済，入らない（アイテム数についての大数の法則）**
- **ずれの中身は，徹底的に単純化した際に落とした社会的要因：アマゾンの恣意性（アルゴリズムの相違，事故・品切），絶版・新刊，大数の法則を越える大規模注文**
- **最小化する評価関数を最小２乗法に変えても**$b < 1$

Amazon.co.jpランキング

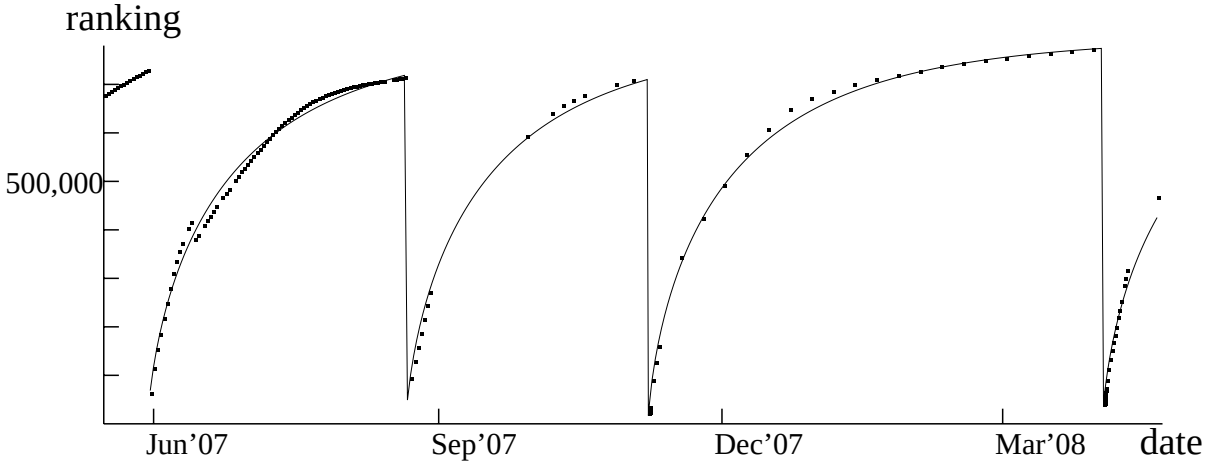


上位への大きな跳び⇔アマゾン書店での注文（1冊でも）　→

確率順位付け模型の結論を当てはめてみる

$O(1$万$)$位は1位とみなす近似で$O(100$万$)$位の現象を理解

アマゾンはロングテールに非ず



3パラメータで98点のデータを当てはめ

$(N^*,\ a^*,\ b^*) = (8 \times 10^5,\ 6 \times 10^{-4},\ 0.81)$

ロングテール型ではなく，ビッグヒット依存型のビジネスモデル

- $b < 1$「不平等」上位$o(N)$冊が売り上げの全てを占める
- 公開情報だけで実証する数理的方法を与えた

３．その他の話題

- 強度が時刻依存性を持つ場合の応用：

2ch.net の書込などの**昼夜差**

- 主定理の応用：

ランキングによる打ち切り判断の合理性 ($b < 1$ の場合)

- 点数の年次変化：

強度の時刻依存性によって原理的には可能

skip

３－１．昼夜差

ジャンプ率に時刻依存性（例：昼夜差）がある場合

$$X_i^{(N)}(t) \sim N - N \int_{L^1_{loc}(\mathbb{R}_+)} e^{-\int_{t_0}^{t} w(s)\,ds} \Lambda(dw)$$

- **共通の時刻依存性の仮定：** $w_i^{(N)}(t) = w_i^{(N)} A(t),\ w_i^{(N)} \geqq 0$

社会活動の昼夜差 は粒子（順位の対象）に無関係と仮定
（差は統計誤差と扱う単純化）

日内変動に影響されないデータ採取計画法

時刻の関数$A(t)$を決めるには**大規模なデータ**が必要

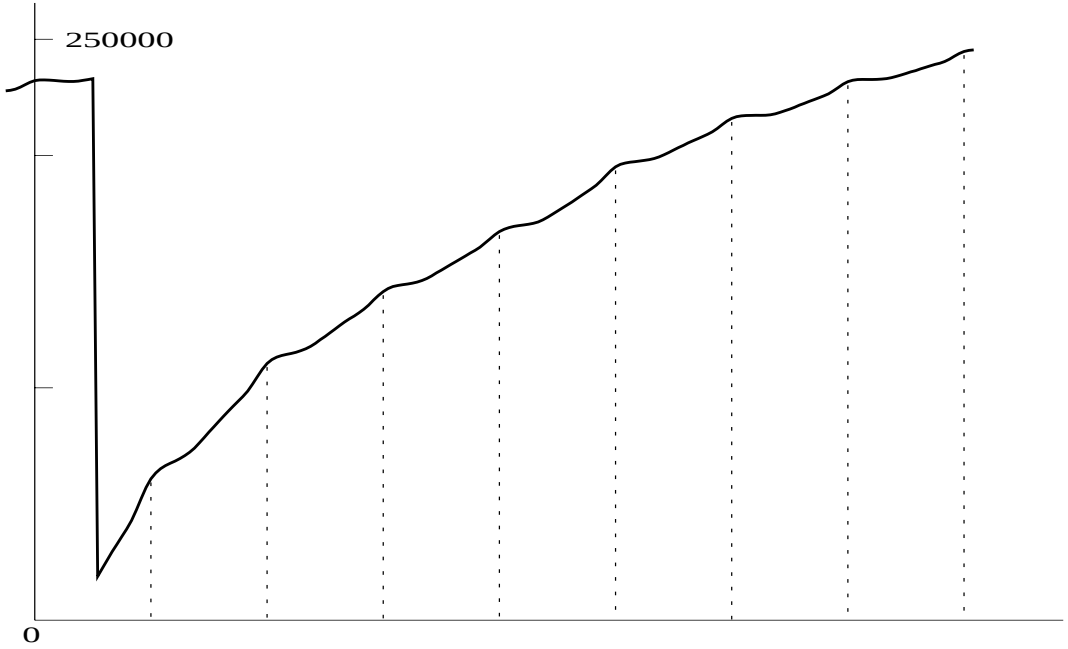
ジャンプ率分布λを先に引き出したいとき

共通時刻依存性＋周期関数（昼夜差）の仮定→$A(t) = t +$ **周期関数**

$$X_i^{(N)}(24n + t_0) \sim N - N \int_{\mathbb{R}_+} e^{-w\, n + a_0} \lambda(dw))$$

- 毎日**定時**に得たデータのみ用る（**逃したら諦める**）場合は，日内変動が無い場合に一致

Amazon.co.jp　社会活動の昼夜差

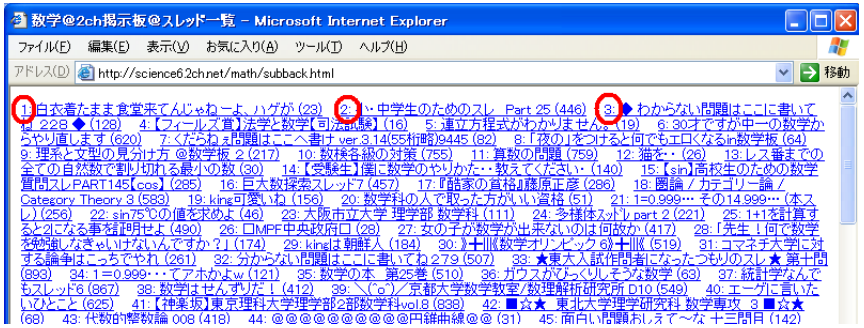


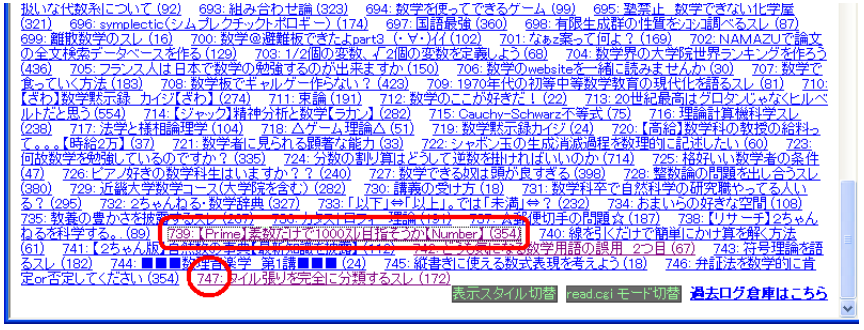
明らかな昼夜差（深夜から早朝まで全体の活動停滞）
→強度の共通時刻依存性を適用
→毎日定時のデータは一様な場合に帰着

- 強度の時刻依存性（昼夜差）のデータからの定量は難しい

アマゾン書店は1時間に一度の更新，web1ページに1点の本の情報

2ch.netのスレッド一覧



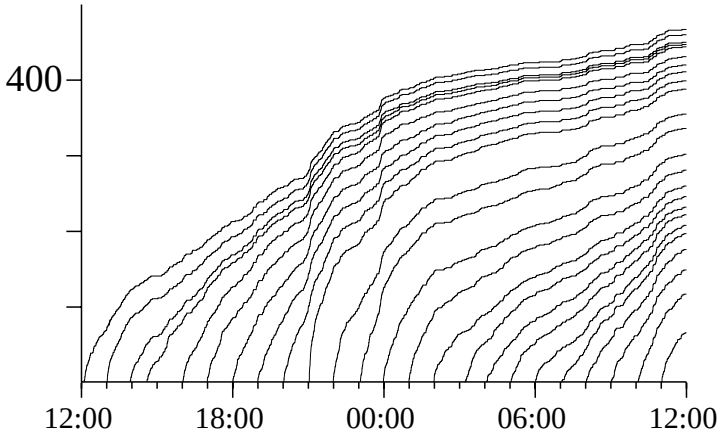


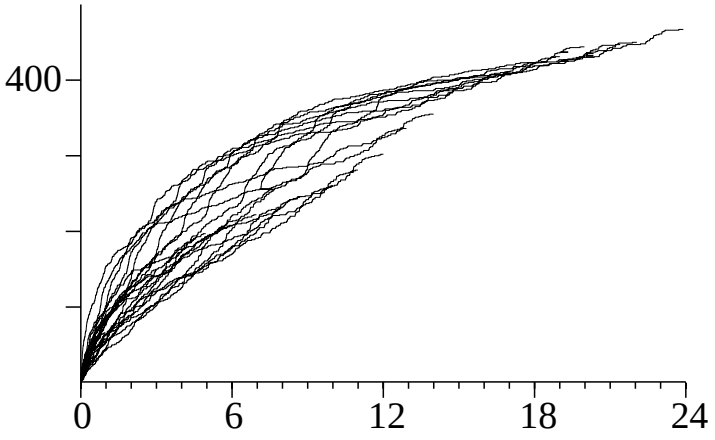
２ちゃんねる：
web**掲示板の巨大な集まり**

スレッド（ページ）一覧：
書き込んだスレッドが１位
move-to-front **規則**
（注：sage**進行は除く）**
N**は**10^3**程度で小さい（ので，数%の揺らぎがある）が，**
先頭に跳ぶ規則が明らかなので，踏み込んだ検証が可能
スレッド一覧の１ページに全スレッドの順位情報→全数調査可能

スレッド一覧の順位変化

強度の日変化の推定





左図は24個のスレッドの，最後に書き込まれて以降の順位変化

右図は最後に書き込まれた時刻を0に取り直して重ねた図

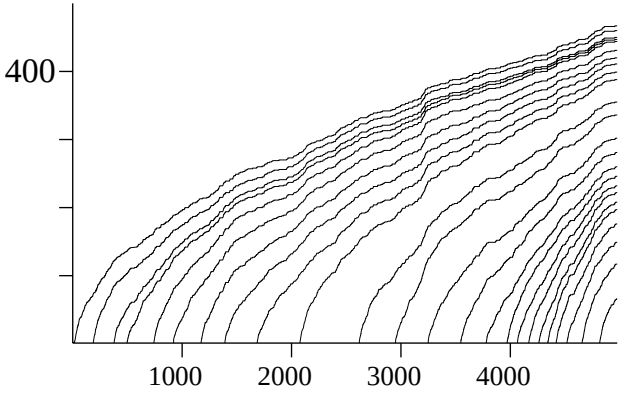
昼夜差のため，1つの関数$y_C(t)$に重ならない

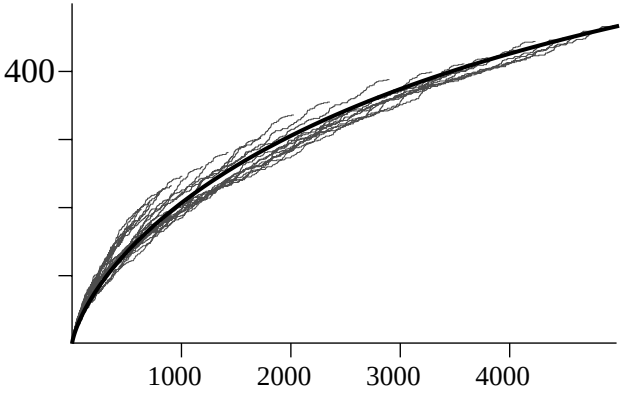
データ自動収集プログラム竹島佑介君（2008年度東北大M2，現富国生命）

集計 小林孝長君（2009年度東北大M2 ，現仙台二高）

共通の時刻依存性の検証とPareto指数の普遍性

(再掲) $X_i^{(N)}(t) \sim N(1 - \int_0^\infty e^{-wS^{(N)}(t)}\lambda(dw));\ \lambda([w,\infty)) = \left(\frac{a}{w}\right)^b,\ w \geqq a$





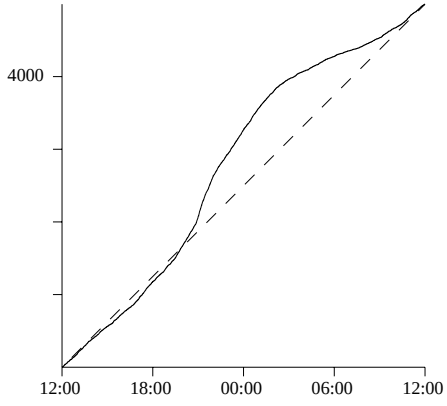
- **共通の時刻依存性は実用的に良さそうである**

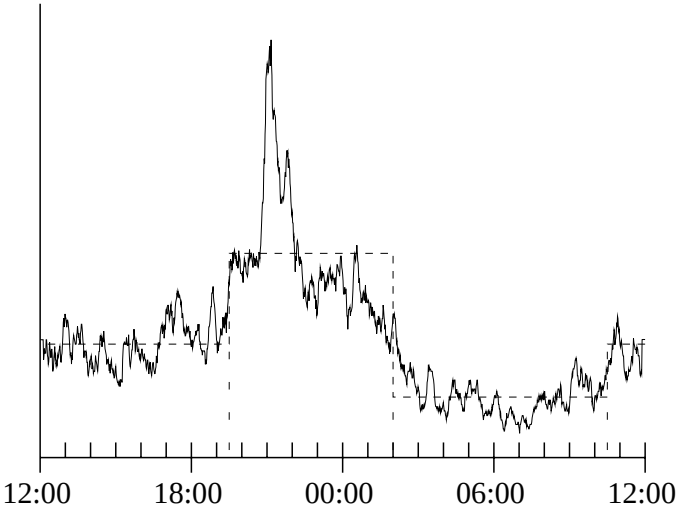
$X_i^{(N)}(A^{-1}(t))$ **(**$A(t) \sim S^{(N)}(t)$**)** → $b = 0.872 < 1$

- Amazon.co.jp **も** 2ch.net **も共通して** $b < 1$

  本の購入はベストセラーに，書き込みは人気スレッドに集中

昼夜差





- 20–01 **時に書込活動活発，**03–09 **時に不活発（２ちゃんねる）（アマゾン書店と整合）**

３－２．主定理とランキングによる打ち切り

時間的に一様，かつ，定常$(t \to \infty)$**な場合，主定理の極限結合分布は**

$$\mu_\infty(dw \times [y,1)) = e^{-wt_0(y)}\lambda(dw);\ 1-y = \int_0^\infty e^{-wt_0(y)}\lambda(dw)$$

$(**)$ **特に，**$t_0(y) > 0,\ y > 0 \quad (\text{cf. } t_0(0) = 0)$

先頭に跳ぶ粒子のうち tail **側** $[y,1)$ **にいたものの割合＝総売上への時刻**t**にランキング下位だった本からの寄与＝**

$$\frac{\int_{\mathbb{R}_+} w\,\mu_\infty(dw \times [y,1))}{\int_{\mathbb{R}_+} w\,\lambda(dw)}$$

$(**)$ → **分子は有限** →

- λ**の平均が有限⇔下位の売上は無視できない** $(b < 1)$
- λ**が平均を持たない⇔下位の売上は無視できる** $(b > 1)$

ロングテール

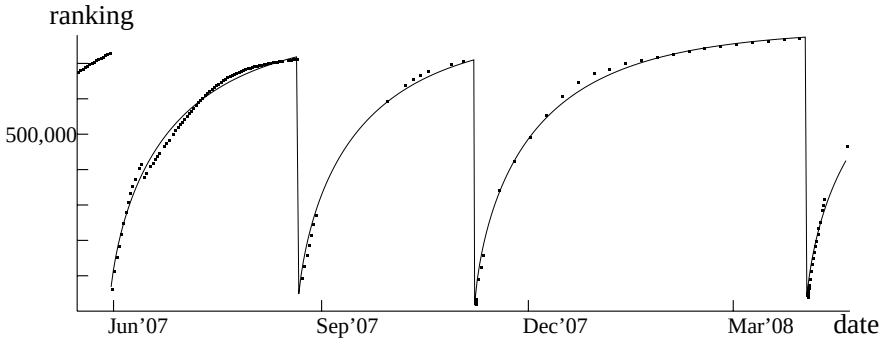
$w_i^{(N)}$ の下位を切るのが正確だろうが，$w_i^{(N)}$ は測定可能量ではない

ある瞬間のランキングの下位を切るとき（機会）損失の見積もり

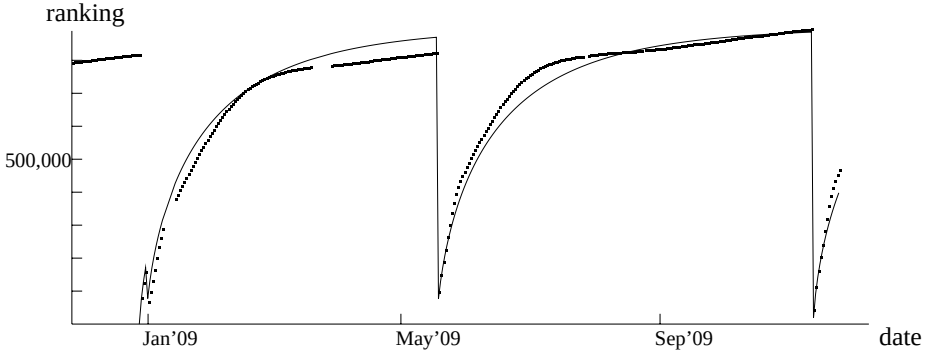
- λ の平均が有限⇔ランキング下位の売上は無視できない　　$(b < 1)$
- λ が平均を持たない⇔下位の売上は無視できる　　$(b > 1)$

（入学試験の原理）

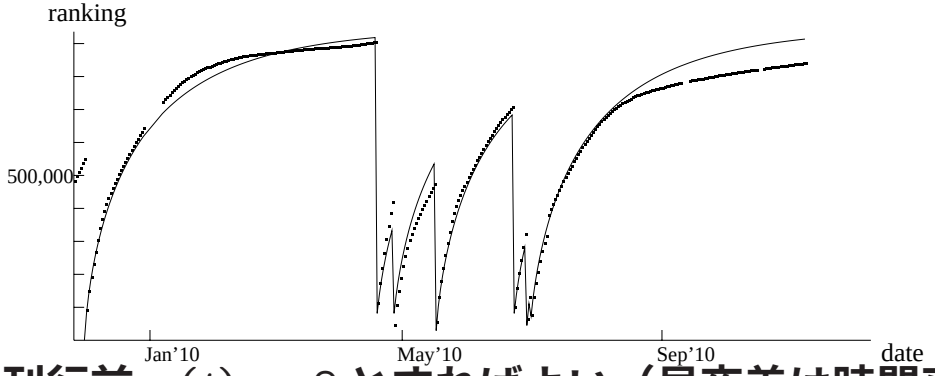
冊数の変化



$N = 80$**万**(2007**年**)



$N = 90$**万**(2009**年**)

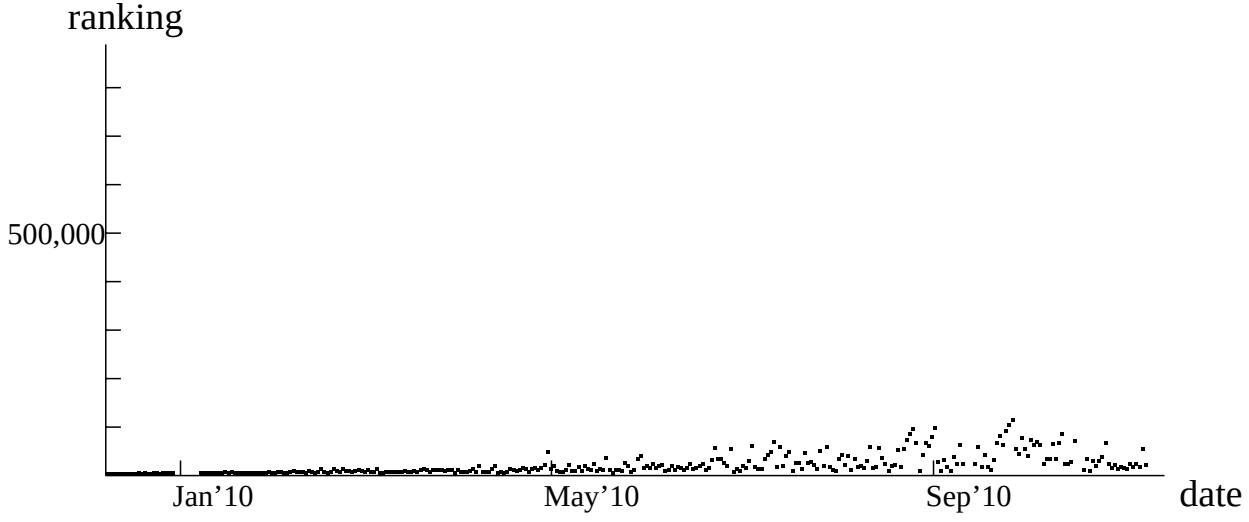


$N = 95$**万**(2010**年**)

刊行前 $w(t) = 0$ **とすればよい（昼夜差は時間変更）．ライフサイクルの検出**

＊．　普遍性とロングテール

アマゾン書店のアルゴリズムは確率順位付け模型か？（多分違う）
なぜ合うか？　普遍性の仮説：
めったに売れない本の流行度は最後に売れた順しかない



左半分（刊行間もなく）１万位未満，右半分７万位未満
ロングテールで問題になる順位は，
常識的に順位を考える領域を１位と同一視する縮尺（新しい現象）skip

確率的な順位付け

下位の$w_i^{(N)}$は測定可能量ではない．

データから決めるには$1/w_i^{(N)} \gg 1$の時間を要する「五十歩百歩」．

順位付けは確率的な概念である．能力の正確な反映ではない．

ロングテールは順位付けが確率的であることが顕著な領域を扱う．

流体力学的極限という一つの帰結を得た．他に何があるか？

文献

服部哲弥,「Amazon ランキングの謎を解く」, 化学同人, 2011.5**.**

K. Hattori, T. Hattori, Stochastic Processes and their Applications **119** (2009) 966–979.

K. Hattori, T. Hattori, Funkcialaj Ekvacioj **52** (2009) 301–319.

K. Hattori, T. Hattori, RIMS Kokyuroku Bessatsu **B21** (2010) 149–162.

Y. Hariya, K. Hattori, T. Hattori, Y. Nagahata, Y. Takeshima, T. Kobayashi, Tohoku Mathematical Journal **63–1** (2011) 77–111.

山下紘史, 2011 **年度研究プロジェクト論文.**

Google**検索キーワード　　服部哲弥**