順序の距離と確率モデル

神嶌 敏弘（産業技術総合研究所）
http://www.kamishima.net/

人工知能学会 第11回 DMSM研究会 (2009.10.18)

「ものさし(尺度)」の種類

長　さ

マイル ➡ キロメートル

1[mile] = 1.609[km]

温　度

華氏 ➡ 摂氏

$$[\text{摂氏}] = \frac{5}{9}([\text{華氏}] - 32)$$

0マイル = 0キロメートル ⇔ 華氏0度 ≠ 摂氏0度

ものさし(尺度) [＝観測→数値の写像] には種類がある

Stevensの尺度水準

二つの量のどのような関係に意味があるかで分類

名義尺度 (nominal scale)

- 一致しているかどうかに意味がある（例：背番号）

順序尺度 (ordinal scale)

- 大小関係に意味がある（例：モースの硬度，段位）

間隔尺度 (interval scale)

- 数値の間隔だけに意味があり，原点はない（例：摂氏，日付）

比率尺度 (ratio scale)

- 原点があり，比率に意味がある（例：重さ，長さ）

順序変量と基本演算

- **演算**

  比較以外の演算である四則演算などはできない

  例：4段と1段の強さを合わせても5段にはならない

- **サンプル集合の代表値**

  加算ができないので，平均値は計算できない

  比較はできるので，中央値や最大値は計算できる

  例：1～3段が一人ずついるとき，平均2段は無意味，中央値が2段は問題ない

- **変換**

  単調関数による変換をしても意味は変わらない

  例：[強]＝log([段]) のような単位を作っても無矛盾

Condorcetのパラドクス

候補：A B C

有権者1	有権者2	有権者3
A ＞ B ＞ C	B ＞ C ＞ A	C ＞ A ＞ B
AとBでは?	BとCでは?	CとAでは?
Aの勝ち	Bの勝ち	Cの勝ち

Condorcetのパラドクス (投票のパラドクス)

個々の有権者の順序関係は明確にもかかわらず,
全体としては順序関係に循環があって勝者が決まらない

※ より一般に順序関係に循環があるをもCondorctのパラドクスと呼ぶこともある

- **Condorcet選択**：一対比較の結果，全ての他の候補に勝る候補がある状態
- **Condorcet勝者**：そのときに勝っている候補

Arrowの不可能性定理

複数の有権者に候補に対する選好順序があるとき，次の四つの公理を満たすように，全体としての選好順序を決めることは不可能

- **[公理 U]**：全ての有権者には，論理的に可能な全ての選好順序をもつことが許される（個人選好の領域無制約性）
- **[公理 P]**：全ての有権者の選好順序で x が y より好まれるなら，全体の選好順序で x が y より好まれる（パレート最適性）
- **[公理 I]**：二つの選好順序の集合で，x と y に関する選好が全ての有権者について一致するなら，それぞれの選好順序に対する全体の選好順序においてx と y の選好は一致（無関係選択肢からの独立性）
- **[公理 ND]**：他の有権者の選好とは無関係に，ある有権者の選好が全体の選好と常に一致してはならない（非独裁）

※どの条件を取り除いても不可能性は解消される

Borda Count

有権者の投票　　A ＞ B ＞ C

1位→3点，2位→2点，3位→1点のポイントを与える

ポイントが最も多い候補が勝者となる

Borda Count は公理Iを満たさない

A＞C＞Bが9人，B＞A＞Cが10人のとき
A＝47p，B＝39p，C＝28pで**Aが勝者**

Cが辞退して，A＞Bが9人，B＞A が10人のとき
A＝28p，B＝29p で**Bが勝者**

AとBの間の選好がCに影響されて公理Iを満たさない!!

順序の表記

順序： $O = x_3 \succ x_1 \succ x_2$ (対象 x_3 が最上位，次が x_1)

順位： 順序中で対象の位置 (例：O 中の x_1 の順位は 2)

順序ベクトル： 第 i 要素は，順位が i の対象 [3, 1, 2]

順位ベクトル： 第 i 要素は，対象 x_{i} の順位 $\begin{bmatrix} 2, & 3, & 1 \\ x_1 & x_2 & x_3 \end{bmatrix}$

群論いうと順位ベクトルや順序ベクトルは対称群（置換群）

全ての可能な長さ m の順位ベクトル：$\mathcal{S}_m$

全ての可能な長さ m の順序ベクトル：$\mathcal{T}_m$

$\pi \in \mathcal{S}_m$ に右から $\xi \in \mathcal{S}_m$ を掛けた $\pi\xi$ は対象番号の付け替え

$\pi \in \mathcal{S}_m$ に左から $\xi \in \mathcal{S}_m$ を掛けた $\xi\pi$ は順位の入れ替え

順序の距離

通常の距離の公理

半正定値性　　対称性　　三角不等式

↓

ラベル不変性 (右不変性)

さらに！　対象の番号を変えても距離は変わらない

$$O_1 = x_3 \succ x_1 \succ x_2$$
$$O_2 = x_2 \succ x_1 \succ x_3$$

$$x_1 \to x'_2$$
$$x_2 \to x'_3$$
$$x_3 \to x'_1$$

$$O'_2 = x'_3 \succ x'_2 \succ x'_1$$
$$O'_1 = x'_1 \succ x'_2 \succ x'_3$$

$$d(O_1, O_2) = d(O'_1, O'_2)$$

群論を使った表記：$d(\pi, \xi) = d(\pi\psi, \xi\psi),\ \pi, \xi, \psi \in \mathcal{S}_m$

順位ベクトルに基づく距離

順序

変換

順位ベクトル

O_1 A > B > C > D

A B C D

1 2 3 4

O_2 D > B > A > C

A B C D

3 2 4 1

通常のベクトルとみなして距離を定義

- Manhattan距離 = Spearman footrule
- Euclid距離
- 2乗Euclid距離 = Spearman距離
- Max距離 Chebyshev距離
- Hamming距離

※Spearman距離は三角不等式を満たさないが便宜的に距離として扱う

Kendall距離とCayley距離

編集距離：一方の順序に，一連の操作を加えてもう一方の順序に変換するとき，その最小変換操作数で距離を定義

Kendall距離：隣接する対象の対を交換する操作

$$A \succ B \succ C \succ D \quad \Rightarrow \quad A \succ C \succ B \succ D$$

Cayley距離：任意の対象の対を交換する操作

$$A \succ B \succ C \succ D \quad \Rightarrow \quad A \succ D \succ C \succ B$$

Kendall距離のもう一つの解釈

A > B > C

B > A > C

A > B　A > C　B > C

NO!　OK　OK

B > A　A > C　B > C

順序対に分解

Kendall距離

逆順の対の数

Ulam距離

Ulam距離の編集操作

誤った位置の対象を取り出す → 正しい位置をスライドさせて空ける → 取り出した対象を正しい位置に配置

B>C>A>D　　B>C　>D　　A>B>C>D

A　　B>C>D　　A

Ulam距離のもう一つの解釈

〈順序の長さ〉－〈最長共通部分列(LCS)の長さ〉

A>B>C>D>E　　D>A>C>B>E

LCS

Ulam距離 ＝ 5 - 3 ＝ 2

順位相関

順位相関係数：順序の距離を [-1,+1] の範囲に正規化

Spearmanの順位相関係数ρ：Spearman距離を正規化

$$\rho = 1 - \frac{6 d_{\mathrm{Spear}}(\pi, \xi)}{m^3 - m}$$

- 二つの順序が独立との帰無仮説の下で

$$\frac{\rho\sqrt{m-2}}{\sqrt{1-\rho^2}} \sim \langle \text{自由度 } m\text{-}2 \text{ のStudentの } t \text{ 分布} \rangle$$

- 二つの順位ベクトルのPearson相関係数に等しい

Kendallの順位相関係数τ：Kendall距離を正規化

$$\tau = 1 - \frac{4 d_{\mathrm{Ken}}(\pi, \xi)}{m(m-1)}$$

- 二つの順序が独立との帰無仮説の下で $\tau \sim \mathcal{N}\left(0, \frac{2(2m+5)}{9m(m-1)}\right)$

距離や順位相関の不等式

Danielsの不等式

$$-1 \leq \frac{3(m+2)}{m-2}\tau - \frac{2(m+1)}{m-2}\rho \leq 1$$

Diaconis-Grahamの不等式

$$d_{\mathrm{Ken}} + d_{\mathrm{Cay}} \leq d_{\mathrm{Foot}} \leq 2d_{\mathrm{Ken}}$$

Durbin-Stuartの不等式

$$d_{\mathrm{Spear}} \geq \frac{4}{3} d_{\mathrm{Ken}} \left(1 + \frac{d_{\mathrm{Ken}}}{m}\right)$$

順序の生成モデル

長さ m の順序は m! 個あるので,
長さの分布は m!-1 パラメータの離散分布

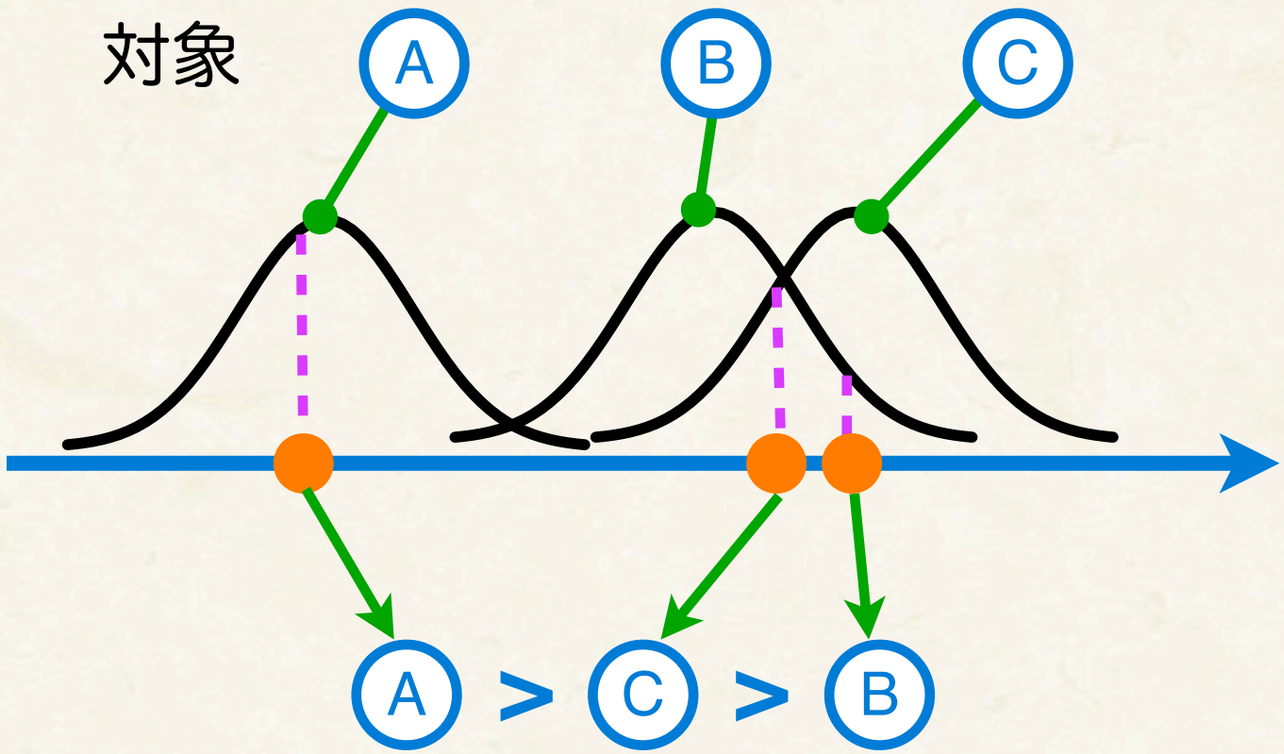
多すぎるので，生成モデルを考慮しつつ，モデルのパラメータを減らす

順序の生成モデルの分類

- **Thurstone型 (Thurstonian)**
  各対象に付随した確率分布に従って発生したスコア順に対象を整列
- **一対比較型 (paired comparison)**
  対ごとに対象を比較し，その比較結果に基づいて全対象を整列
- **距離ベース型 (distance-based)**
  モード順位付けからの距離で決まる確率分布に従って順序を生成
- **多段階型 (multistage)**
  先頭から末尾に向かって対象を逐次的に整列

Thurstone型

Thurstone型 (Thurstonian) / 順序統計量型 (order statistics)



各対象ごとに，それに付随した確率分布に従ってスコアをサンプリング

サンプリングしたスコアの順に対象を整列する

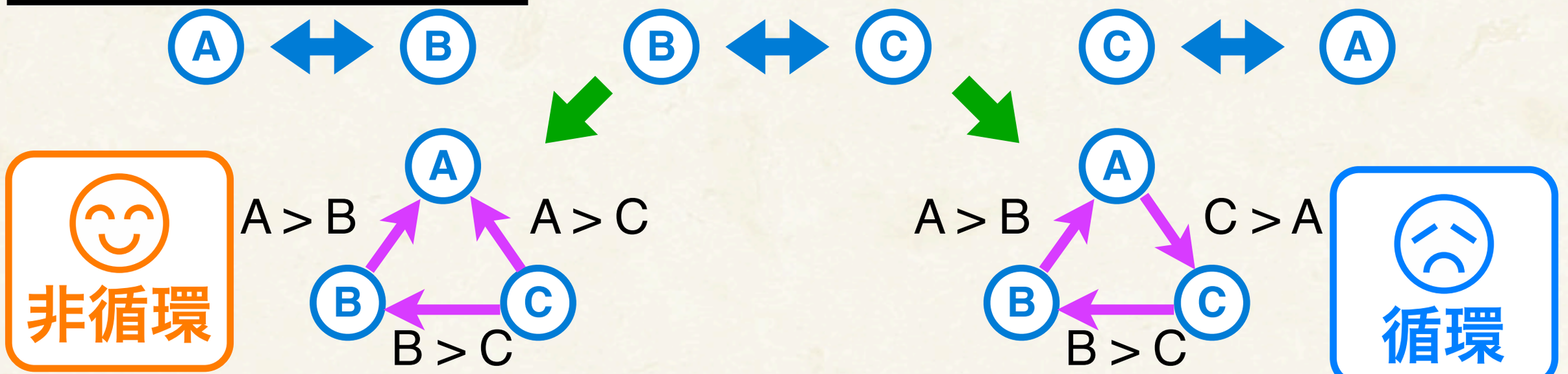
スコアの分布

- **正規分布：**Thurstoneの一対比較の法則
- **Gumbel分布：**累積密度分布が $1 - \exp(-\exp((x_i - \mu_i)/\sigma)$

一対比較型

一対比較型 (paired comparison)

対象を対ごとに比較



順序 A > B > C を生成

廃棄して再試行

パラメータ化

- **Babinton Smithモデル:** ${}_nC_2$ 個のパラメータがある飽和モデル
- **Bradley-Terryモデル:** $\Pr[x_i \succ x_j] = \frac{v_i}{v_i + v_j}$

距離ベース型

距離ベース型 (distance-based)

集中度パラメータ　　モード順序／モード順位付け

$$\Pr[O] = C(\lambda)\exp(-\lambda d(O, O_0))$$

正規化定数　　距離

距離

- **Spearman距離：** Mallowsの θ モデル
- **Kendall距離：**Mallowsの φ モデル

これらは，次式で定義される一対比較モデルであるMallowsもでるにおいて，パラメータが ϕ=1 や θ=1 である特殊な場合に該当

$$\Pr[x_i \succ x_j] = \frac{\theta^{i-j}\phi^{-1}}{\theta^{i-j}\phi^{-1}+\theta^{j-i}\phi}$$

多段階型 (Plackett-Luce)

Plackett-Luceモデル

ある対象を次に整列する確率は，その対象のパラメータを，まだ整列していない対象のパラメータの総和で割ったもの

例：対象 {A,B,C,D} を A＞C＞D＞B の順序に整列

$$\Pr[A] = \frac{\theta_A}{\theta_A + \theta_B + \theta_C + \theta_D}$$

θ_A ― 最上位対象のパラメータ

分母 ― 全対象のパラメータの総和

$$\Pr[A>C \mid A] = \frac{\theta_C}{\theta_B + \theta_C + \theta_D}$$

θ_C ― 第2位の対象のパラメータ

分母 ― A以外の対象のパラメータの総和

$$\Pr[A>C>D \mid A>C] = \frac{\theta_D}{\theta_B + \theta_D}$$

$$\Pr[A>C>D>B \mid A>C>D] = \theta_B / \theta_B = 1$$

順序 A > C > D > B が生成される確率は

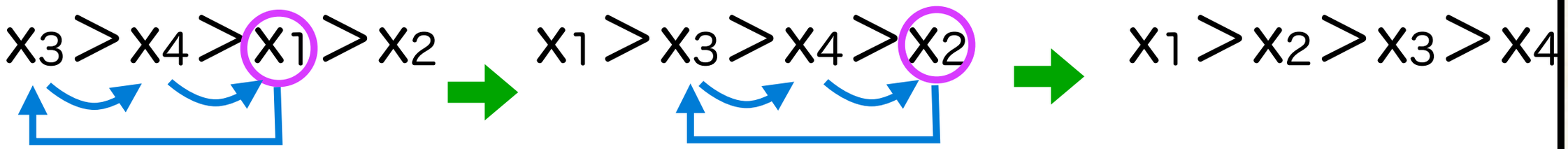
$$\Pr[A>C>D>B] = \Pr[A]\ \Pr[A>C \mid A]\ \Pr[A>C>D \mid A>C]\ 1$$

多段階型 (φ-component)

φ-componentモデル (一般化Mallowsφモデル)

Ulam距離風に対象を移動させたときに，スライドした対象数を考える

例：$x_3 > x_4 > x_1 > x_2$ → $x_1 > x_2 > x_3 > x_4$



$$\Pr[O] = \prod_{i=1}^{m-1} \exp(\theta_i u_i - \phi(\theta_i; m-i+1))$$

$$\phi(\theta; q) = \log\left(\frac{1-\exp(\theta q)}{q(1-\exp(\theta))}\right)$$

全てのパラメータθ_iが等しいとMallowsφモデルと等価

不完全順序

全ての対象に順序関係があった➡完全順序 (complete ranking)

不完全順序 (incomplete ranking)：許される完全順序の集合

同順位 $\{x_1, x_2, x_3\}$ のうち x_3 は他より優れるが，x_1 と x_2 は同等

部分集合 $\{x_1, x_2, x_3\}$ のうち，x_1, x_2 には $x_1 > x_2$ の順序関係が与えられているが，x_3 については全く不明

⬇

条件に無矛盾な全ての完全順序の集合で表現

同順位の例：x_3 が x_1 と x_2 の両方より上位という条件を満たす完全順序の集合 $\{ x_3 > x_2 > x_1, x_3 > x_1 > x_2 \}$

部分集合の例：x_3 は $x_1 > x_2$ 中のどの順位にもなりうるので，不完全順序は集合 $\{ x_3 > x_1 > x_2, x_1 > x_3 > x_2, x_1 > x_2 > x_3 \}$

Hausdorff距離

不完全順序間の距離＝順序の集合間の距離

集合の要素間の平均距離を考えてみる

不完全順序 R と Q の間の平均距離

$$d(R,Q) = \frac{1}{|R||Q|} \sum_{\pi \in R} \sum_{\xi \in Q} d(\pi, \xi)$$

d(R, R)＝0 を満たさない！

↓

Hausdorff距離

$$d(R,Q) = \max\{\max_{\pi \in R} \min_{\xi \in Q} d(\pi, \xi), \max_{\xi \in Q} \min_{\pi \in R} d(\pi, \xi)\}$$

群論を使う方法

top-k 順序：上位 k 個の対象は分かっているが，残りは未知

特定の順位だけ分かっている不完全順序は部分群を使って表現可能

例：{x1…x5} で1位と2位だけ分かっている

S_{n-k}={ [1,2,3,4,5], [1,2,3,5,4], [1,2,4,3,5]…[1,2,5,4,3] }

紫の部分だけ固定した部分群を導入する．ある順位ベクトル π に左からこの部分群を掛けた $S_{n-k}\pi$ は，1位と2位がπと同じで残りは全てのパターンを含むような不完全順序になる．

部分群による不完全順序間のHausdorff距離は効率的に計算可能

$$d_{\mathrm{Ken}}(S_{m-k}\pi, S_{m-k}\xi) = d_{\mathrm{Ken}}(A) + h(m+k-\frac{h-1}{2}) - \sum_{i\in B}\pi(i) - \sum_{i\in C}\xi(i)$$

$$d_{\mathrm{Spear}}(S_{m-k}\pi, S_{m-k}\xi) = \sum_{i\in A}(\pi(i)-\xi(i))^2 + h^2(m-k-h)$$
$$+\max\Big\{\sum_{j=1}^{h}(m+1-j-\pi_B(j))^2+\sum_{j=1}^{h}(k+j-\xi_C(j))^2, \sum_{j=1}^{h}(k+j-\pi_B(j))^2+\sum_{j=1}^{h}(m+1-j-\xi_C(j))^2\Big\}$$

※ 記号の詳細などは予稿を参照のこと

midrank

midrank：同順位の対象集合に，それらの対象が取り得る順位の平均順位を割り当てる

例：$x_1 > x_2 \sim x_3 > x_4$ の場合

不完全順序は { $x_1 > x_2 > x_3 > x_4$, $x_1 > x_3 > x_2 > x_4$ }

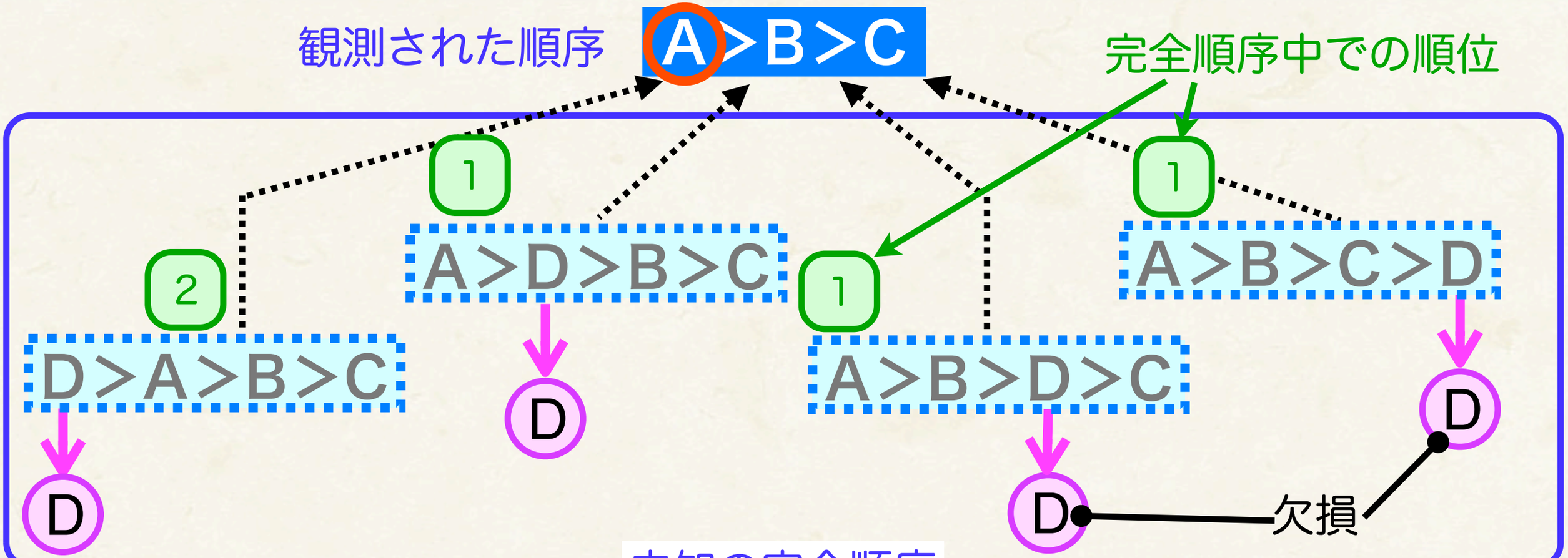
x2 と x3 は共に 2位 か 3位 になる

x2 と x3 に共にmidrank 2.5 を与える

順位ベクトルで書くと [1, 2.5, 2.5, 4] となる

強引に見えるが，同順位のある順位相関などはこうして定義される

期待順位



元の順序は等確率(=1/4)でこれらのうちのいずれか

$$\text{期待順位} = \frac{\langle\text{完全順序の長さ}\rangle + 1}{\langle\text{観測順序の長さ}\rangle + 1} \times \langle\text{観測順序中の順位}\rangle$$

$$\text{対象Aの期待順位} = 1 \times 2 \times \frac{1}{4} + 3 \times 1 \times \frac{1}{4} = \frac{4+1}{3+1} \times 1 = \frac{5}{4}$$

まとめ

- **順序変量：**大小関係だけに意味がある
- **順序間の距離**
  Spearman footrule, Spearman距離，Kendall距離，Cayley距離，Ulam距離など
- **順序の確率分布**
  - **Thurstone型：**Thurstoneの一対比較の法則，Gumbel分布
  - **一対比較型：**Babington-Smithモデル，Bradley-Terryモデル，Mallowsモデル
  - **距離ベース型：**Mallows ϕ モデル，Mallows θ モデル
  - **多段階型：**Plackett-Luceモデル，ϕ-componentモデル

寿司データ：順位法と採点法による寿司の嗜好に関するデータ
http://www.kamishima.net/sushi/