

生物多様性と群集動態：定量化の数理と統計的推定法

11月8日

14:00-14:30 島谷健一郎(統数研) 生物多様性の統計数理の曼陀羅

14:30-14:40 開催地からのあいさつ＋事務連絡 田中健太(筑波大菅平)

14:40-15:50 深谷肇一(環境研) 長期・広域の生態系モニタリングのモデル化：種多様性の変化を推測する

16:00-17:30 長田穰(東北大) 複雑に相互作用する生物群集と平均場近似

11月9日

9:00-10:30 東樹宏樹(京都大) 大規模データで群集集合の共通原理を探る

10:40-12:10 川津一隆(東北大) ランダム行列の観点から解き明かす生態系の創発特性

13:00-14:30 エクスカーション 菅平高原

15:00-15:50 田中健太(筑波大) 草原の継続期間による植物・昆虫群集の遷移：多様性と種特性の変化

15:50-16:40 門脇浩明(京都大) 理論・統計・シミュレーションの三位一体～あたらしい生態学教育をめざして～

16:40-17:10 島谷健一郎(統数研) 生物多様性指数の統計数理 サンプル被覆度とデータからの推定法

11月10日

9:00-10:30 近藤倫生(東北大) 種間相互作用とは何か：生物学的レベルの重要性

10:30-11:10 島谷健一郎(統数研) 生物多様性指数の統計数理

11:10 - 12:00 総合討論

- ・Chao以外の観察されなかった種数推定法
- ・ヒル数以外の多様性指数とその応用例

統計数理研究所公開講座 2023年6月13日

行列の固有値という数学： 統計学と生態学からの再入門

島谷健一郎(統数研)

1. 固有値をどう習ったか。線形代数学における固有値という概念
2. 生態学における個体群成長率という固有値の解釈1
3. 生態学における個体群成長率という固有値の解釈2
4. 主成分分析と固有値
5. 定数係数線形微分方程式の解の挙動と固有値
6. 直観でとらえられる固有値、固有値の教育法雑感

定数係数線形微分方程式

$$\frac{dx(t)}{dt} = rx(t) \quad x(t) = e^{rt} x_0 \quad x_0: \text{初期値}$$

連立定数係数線形微分方程式(2変数)

実際問題、連立の意味がない場合

あえてベクトルと行列で書くと

$$\begin{cases} \frac{dx}{dt} = ax \\ \frac{dy}{dt} = by \end{cases} \quad \begin{cases} x = e^{at} x_0 \\ y = e^{bt} y_0 \end{cases} \quad \begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} e^{at} x_0 \\ e^{bt} y_0 \end{pmatrix}$$

(本当の)連立定数係数線形微分方程式(2変数)

$$\begin{cases} \frac{dx}{dt} = y \\ \frac{dy}{dt} = -x \end{cases} \quad \text{ベクトルと行列で書くと} \quad \frac{d}{dt} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \text{と置いて} \quad \frac{d\mathbf{x}}{dt} = \mathbf{M}\mathbf{x}$$

一般に高次の定数係数線形微分方程式は、1次元のときのスカラーをベクトルと行列に換えれば同じように解ける

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}\mathbf{x} \quad \text{の解は} \quad \mathbf{x}(t) = e^{\mathbf{M}t} \mathbf{x}_0 \quad \mathbf{x}_0: \text{初期値(初期ベクトル)}$$

行列の指数関数
定義

$$e^{\mathbf{M}t} = \mathbf{E} + t\mathbf{M} + \frac{t^2}{2!}\mathbf{M}^2 + \frac{t^3}{3!}\mathbf{M}^3 + \dots \quad \mathbf{E}: \text{単位行列}$$

以下のような命題が知られています

- ・全成分が収束する
- ・行列 \mathbf{A} と \mathbf{B} が可換 ($\mathbf{AB} = \mathbf{BA}$) なら $e^{\mathbf{A}} e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$
- ・ $\mathbf{P}^{-1} e^{\mathbf{M}} \mathbf{P} = e^{\mathbf{P}^{-1}\mathbf{M}\mathbf{P}} \quad \because (\mathbf{P}^{-1}\mathbf{M}\mathbf{P})^2 = \mathbf{P}^{-1}\mathbf{M}\mathbf{P}\mathbf{P}^{-1}\mathbf{M}\mathbf{P} = \mathbf{P}^{-1}\mathbf{M}\mathbf{P}$
- ・ ...
などからわかる

一般に高次の定数係数線形微分方程式は、1次元のときのスカラーをベクトルと行列に換えれば同じように解ける

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}\mathbf{x} \quad \text{の解は} \quad \mathbf{x}(t) = e^{\mathbf{M}t} \mathbf{x}_0 \quad \mathbf{x}_0: \text{初期値(初期ベクトル)}$$

証明: 行列の指数関数の定義を使う、解の一意性を使う、等々

行列の指数関数の計算
対角なら自明

$$\mathbf{M} = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{pmatrix} \quad e^{\mathbf{M}} = \begin{pmatrix} e^{a_1} & 0 & \cdots & 0 \\ 0 & e^{a_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{a_n} \end{pmatrix}$$

\mathbf{M} が行列 \mathbf{T} で対角化可能 ($\mathbf{T}^{-1}\mathbf{M}\mathbf{T} = \mathbf{D}$, $\mathbf{D} = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{pmatrix}$, a_1, \dots, a_n は \mathbf{M} の固有値) なら、

$$\mathbf{x}(t) = \mathbf{T} e^{\mathbf{T}^{-1}\mathbf{M}\mathbf{T}t} \mathbf{T}^{-1} = \mathbf{T} e^{\mathbf{D}t} \mathbf{T}^{-1}$$

一般に高次の定数係数線形微分方程式は、1次元のときのスカラーをベクトルと行列に換えれば同じように解ける

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}\mathbf{x} \quad \text{の解は} \quad \mathbf{x}(t) = e^{\mathbf{M}t} \mathbf{x}_0 \quad \mathbf{x}_0: \text{初期値(初期ベクトル)}$$

対角化できない場合、ジョルダン標準形に直すことで、指数関数と多項式の積の1次結合で解は書けることが示されている

$$\begin{pmatrix} a & 1 & 0 & 0 & \cdots & 0 \\ 0 & a & 1 & 0 & \cdots & 0 \\ 0 & 0 & a & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a & 1 \\ 0 & 0 & 0 & \cdots & 0 & a \end{pmatrix}$$

a : 重根の固有値
 行列サイズ＝重根の重複度
 異なる重根固有値の数だけこのようなブロックが並ぶ(単根固有値は対角成分のみ)

M が行列 T で対角化可能($T^{-1}MT = D, D = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{pmatrix}$, a_1, \dots, a_n は M の固有値)なら、

$$\mathbf{x}(t) = \mathbf{T} e^{\mathbf{T}^{-1}\mathbf{M}\mathbf{T}t} \mathbf{T}^{-1} = \mathbf{T} e^{\mathbf{D}t} \mathbf{T}^{-1}$$

定数係数線形微分方程式 $\frac{d\mathbf{x}}{dt} = \mathbf{M}\mathbf{x}$ の解は

$$\mathbf{x}(t) = \mathbf{T}e^{\mathbf{D}t}\mathbf{T}^{-1} \quad \mathbf{D} = \mathbf{T}^{-1}\mathbf{M}\mathbf{T} = \begin{pmatrix} a_1 & * \\ 0 & a_2 \end{pmatrix} \quad a_1, a_2 \text{ は } \mathbf{M} \text{ の固有値}$$

平衡点: $\mathbf{x}(t) = \mathbf{C}$ (\mathbf{C} は定数ベクトル) が解のとき、 \mathbf{C} を平衡点という

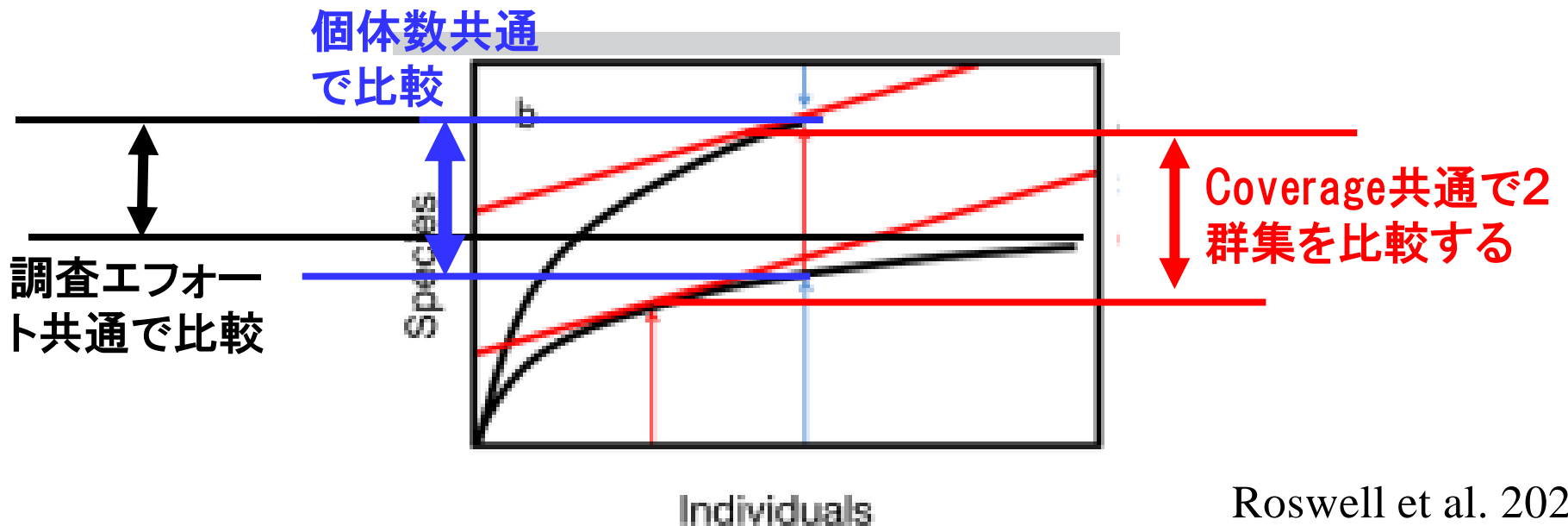
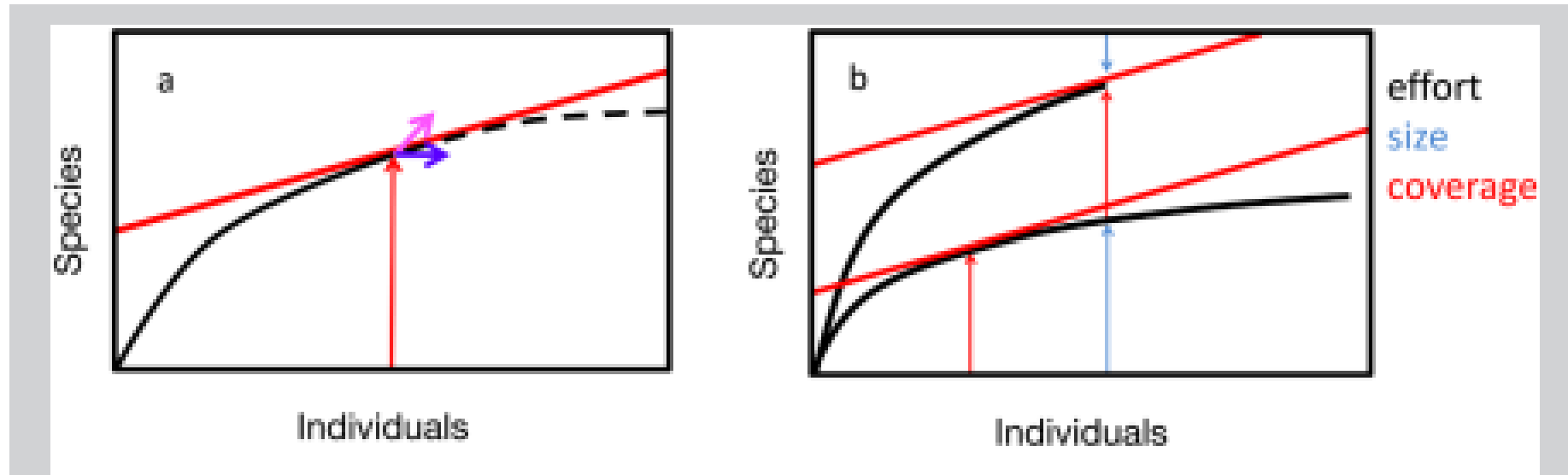
$\frac{d\mathbf{x}}{dt} = \mathbf{M}\mathbf{x}$ では、 $\mathbf{C} = \mathbf{0}$ は平衡点

平衡点付近での解の挙動は \mathbf{M} の固有値で分類できる

対角化可能 $\left\{ \begin{array}{l} a_1, a_2 \text{ は実の単根} \left\{ \begin{array}{l} \text{同符号} \begin{cases} a_1, a_2 < 0 \\ a_1, a_2 > 0 \end{cases} \\ \text{異符号} (a_1 a_2 < 0) \end{array} \right. \\ \text{実の重根} \\ \text{複素数の2根} \left\{ \begin{array}{l} \text{実部} > 0 \\ \text{実部} = 0 \\ \text{実部} < 0 \end{array} \right. \end{array} \right.$

対角化できない(必然的に重根)

サンプル被覆度の応用例: 2つの群集の比較の基準



サンプル被覆度: 手にしたデータで集団の何パーセントを抑えられたか(データごとに変動する確率変数)

定義

サンプル数 n のデータにおけるサンプル被覆度 (sample coverage)

$$C_n = \sum_{i=1}^S p_i \underbrace{I(X_i > 0)}_{\text{データによって変動}} = 1 - \sum_{i=1}^S p_i I(X_i = 0)$$

未知

$I(\cdot)$: indicator function
= 1 ()内が正しいとき
= 0 otherwise

A_k : k 個体観察された種の相対頻度の平均値(データごとに変動する確率変数)

$$A_k = \frac{\sum_{i=1}^S p_i I(X_i = k)}{\sum_{i=1}^S I(X_i = k)}$$

観察された個体数が k である種の数 $F_k = \sum_{i=1}^S I(X_i = k)$

データ(確率変数の実現)が与えられたら小文字

$$a_k = \frac{\sum_{i=1}^S p_i I(x_i = k)}{f_k}$$

$C_n = 1 - A_0 F_0$ とも書ける

サンプル被覆度の期待値 $E[C_n] = 1 - E[\sum_{i=1}^S p_i I(X_i = 0)] = 1 - \sum_{i=1}^S p_i (1 - p_i)^n$

サンプル被覆度 $C_n = 1 - A_0 F_0$ の期待値をデータから推定するため、 A_k と F_k の漸化式を作ります。まず、その準備のための式変形をしておきます。

$$\begin{aligned} \mathbf{E}\left[\sum_{i=1}^S \frac{p_i}{1-p_i} I(X_i = k)\right] &= \sum_{i=1}^S \frac{p_i}{1-p_i} \mathbf{E}[I(X_i = k)] && \text{確率変数の期待値} \\ &= \sum_{i=1}^S \frac{p_i}{1-p_i} \binom{n}{k} p_i^k (1-p_i)^{n-k} = \sum_{i=1}^S \frac{n(n-1)\cdots(n-k+1)}{k!} p_i^{k+1} (1-p_i)^{n-k-1} \end{aligned}$$

分子分母に $k+1$ と $n-k$ をかけると、

$$= \sum_{i=1}^S \frac{k+1}{n-k} \cdot \frac{n(n-1)\cdots(n-k+1)(n-k)}{(k+1)!} p_i^{k+1} (1-p_i)^{n-k-1} = \frac{k+1}{n-k} \sum_{i=1}^S \binom{n}{k+1} p_i^{k+1} (1-p_i)^{n-k-1}$$

ところで、

$$\mathbf{E}[F_k] = \sum_{i=1}^S \mathbf{E}[I(X_i = k)] = \sum_{i=1}^S (P(X_i = k) \cdot 1 + P(X_i \neq k) \cdot 0) = \sum_{i=1}^S \binom{n}{k} p_i^k (1-p_i)^{n-k}$$

でした。上式右辺に代入すると、

$$\mathbf{E}\left[\sum_{i=1}^S \frac{p_i}{1-p_i} I(X_i = k)\right] = \frac{k+1}{n-k} \mathbf{E}[F_{k+1}]$$

確率変数 X_i (種 i の出現個体数) のひとつの実現を与える。データは期待値にある程度は近いはずなので、右辺は $\frac{r+1}{n-r} \cdot f_{k+1}$ と近い。

左辺にある p_i たち中の、個体数がちょうど k 個となった p_i の平均が a_k だった。だから、ひとつの実現を与えると p_i たちはどれも a_k に近い値になる。 $\sum_{i=1}^S I(X_i=k)$ は f_k となる。従って、左辺は $\frac{a_k}{1-a_k} \cdot f_k$ に近い

$$\frac{r+1}{n-r} \cdot f_{k+1} \approx \frac{a_k}{1-a_k} \cdot f_k$$

a_k について解くことで、 a_k の推定量を得る $\hat{a}_k = \frac{(r+1)f_{k+1}}{(n-r)f_k + (r+1)f_{k+1}}$

$k=0$ とすると $\hat{a}_0 = \frac{f_1}{nf_0 + f_1}$ $\hat{f}_0 = \frac{n-1}{n} \cdot \frac{f_1^2}{2f_2}$ (Chao の推定量) を代入して

$$\hat{a}_0 = \frac{2f_2}{(n-1)f_1 + 2f_2}$$

従って

$$\hat{a}_0 \hat{f}_0 = \frac{2f_2}{(n-1)f_1 + 2f_2} \cdot \frac{n-1}{n} \cdot \frac{f_1^2}{2f_2} = \frac{f_1}{n} \cdot \frac{(n-1)f_1}{(n-1)f_1 + 2f_2}$$

$C_n = 1 - A_0 F_0$ に推定量 $\hat{a}_0 \hat{f}_0 = \frac{f_1}{n} \cdot \frac{(n-1)f_1}{(n-1)f_1 + 2f_2}$ を代入して

$$\hat{C}_n = 1 - \frac{f_1}{n} \cdot \frac{(n-1)f_1}{(n-1)f_1 + 2f_2}$$

n が大きいと後半はほぼ1だから $\hat{a}_0 \hat{f}_0 \approx \frac{f_1}{n}$ という式も得られる

0個体観察された種が群集全体で占める割合は、1個体だけ観察された種数 f_1 で推定できる（何か不思議...、でもそうなんです）

では、1個体だけ観察された種が占める割合はどう推定するかとい

うと、式 $\hat{a}_k = \frac{(r+1)f_{k+1}}{(n-r)f_k + (r+1)f_{k+1}}$ で $k=1$ とし、両辺に f_1 をかけて

$$\hat{a}_1 f_1 = \frac{2f_1 f_2}{(n-1)f_1 + 2f_2}$$

1個体観察された種が群集全体で占める割合は、1個体だけ観察された種数 f_1 と2個体だけ観察された種数 f_2 から推定できる

データからヒル数を推定する話へ進みます

ヒル数

Jost (2006), Hill (1973)

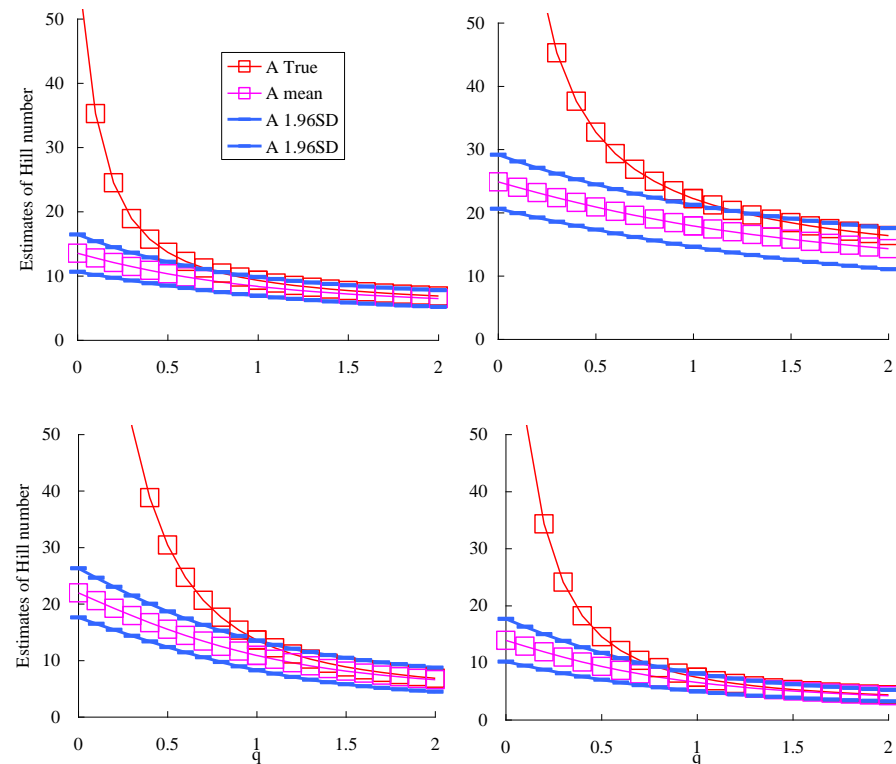
$${}^qD = \left(\sum_{s=1}^S p_s^q \right)^{\frac{1}{1-q}} \quad (q \geq 0)$$

$q = 0$ のとき 0D は種数 S

$q = 2$ のとき ジニ・シン普森指数 (の逆数)

$q = 1$ のとき シャノン指数 (を e の肩)

ヒル数の式に観察された
相対頻度を代入すると
過少推定になる



種数－個体数曲線:横軸に個体数 k 、縦軸にそのときに観察される種数 $S(k)$ をプロットしたものを順に結んだ曲線(折れ線?)

種数－個体数曲線の傾き $\Delta(k) = \frac{S(k+1) - S(k)}{(k+1) - k} = S(k+1) - S(k)$

$S(k)$ はサンプルに依存して変動する確率変数, 傾き $\Delta(k)$ も確率変数である。それらの期待値を考える

k 個体の中に種 i が1度も観察されない確率 $= (1 - p_i)^k$

1度は観察される確率 $= 1 - (1 - p_i)^k$

どの種も、観察されれば1、観察されなければ0と数えられるから、 $S(k)$ の期待値は

$$\mathbf{E}[S(k)] = \sum_{i=1}^S \{ (1 - (1 - p_i)^k) \cdot 1 + (1 - p_i)^k \cdot 0 \} = \sum_{i=1}^S 1 + \sum_{i=1}^S (1 - p_i)^k = S - \sum_{i=1}^S (1 - p_i)^k$$

同様に、

$$\mathbf{E}[S(k+1)] = S - \sum_{i=1}^S (1 - p_i)^{k+1} = S - \sum_{i=1}^S (1 - p_i)(1 - p_i)^k = S - \sum_{i=1}^S (1 - p_i)^k + \sum_{i=1}^S p_i(1 - p_i)^k$$

下から上を引いた残りが $\Delta(k)$ の期待値

$$\mathbf{E}[\Delta(k)] = \mathbf{E}[S(k+1)] - \mathbf{E}[S(k)] = \sum_{i=1}^S p_i (1 - p_i)^k$$

$S(0) = 0, S(1) = 1$ だから、常に $\Delta(1) = 1$

注) サンプル被覆度との関係 $\mathbf{E}[C_n] = 1 - \sum_{i=1}^S p_i (1 - p_i)^n = 1 - \mathbf{E}[\Delta(n)]$

ヒル数 ${}^q D = \left(\sum_{i=1}^S p_i^q \right)^{\frac{1}{1-q}}$ のべき乗の中を D_q とする $D_q = \sum_{i=1}^S p_i^q$ (${}^q D = D_q^{1/1-q}$)

$D_q = \sum_{i=1}^S p_i^q = \sum_{i=1}^S p_i (1 - (1 - p_i))^{q-1}$ と、わざとシグマの中を長くする

2項分布を、べきを自然数から一般の実数
べきへ拡張した関数のテーラー展開 $(x + y)^m = \sum_{k=0}^{\infty} \binom{m}{k} x^k y^{m-k}$

$y = 1, x$ を $-x$ にした $(1 - x)^m = \sum_{k=0}^{\infty} \binom{m}{k} (-1)^k x^k$ を使う (m には $q - 1$ を代入)

$$D_q = \sum_{i=1}^S p_i \sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k (1 - p_i)^k = \sum_i \sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \mathbf{E}[\Delta(k)]$$

次の等式が成り立ちます(証明は数学としてテクニカルなので後回し(たぶん省略))

$$\mathbf{E}\left[\frac{X_i}{n} \frac{\binom{n-X_i}{k}}{\binom{n-1}{k}}\right] = p_i(1-p_i)^k = \mathbf{E}[\Delta(k)] \quad (k < n)$$

$$D_q = \sum_{k=0}^{n-1} \binom{q-1}{k} (-1)^k \mathbf{E}[\Delta(k)] + \sum_{k=n}^{\infty} \binom{q-1}{k} (-1)^k \mathbf{E}[\Delta(k)]$$

の第1項の $\mathbf{E}[\Delta(k)]$ には、上式左辺の確率変数の実現値を代入した式

$$\hat{\Lambda}(k) = \sum_{i=1}^s \frac{x_i}{n} \frac{\binom{n-x_i}{k}}{\binom{n-1}{k}} \quad (k < n)$$

を使う

x_s : 種 s の観察個体数

第2項(無限和)

サンプル数 n のデータがあるとき、そこからさらにサンプル数 m のデータを追加したときに新たに観察される種数を考える。

$S(n+m)$ という確率変数をサンプル数 n のデータという条件の元で考えることを意味する。

観察されていない種の平均相対頻度は a_0 だから、それが新たな m 個のサンプルで少なくとも1回出てくる確率は

$$1 - (1 - a_0)^m$$

こうした種は全部で f_0 種あるから、新たな m 個で出てくる新しい種の数
の期待値は

$$f_0 (1 - (1 - a_0)^m)$$

サンプル数 n のデータ (x_1, \dots, x_S) の元での $S(n + m)$ の条件付き期待値

サンプル数 n の中で観察された種数を s_{obs} と書くと(実際の数値なので
小文字の s)、

$$\mathbf{E}[S(n + m) \mid (x_1, \dots, x_S)] = s_{obs} + f_0 (1 - (1 - a_0)^m)$$

同様にして

$$\mathbf{E}[S(n + m + 1) \mid (x_1, \dots, x_S)] = s_{obs} + f_0 (1 - (1 - a_0)^{m+1})$$

$\Delta(n + m) = S(n + m + 1) - S(n + m)$ という確率変数の条件付き期待値
は下から上を引いて

$$\mathbf{E}[\Delta(n+m) | (x_1, \dots, x_S)] = f_0(1 - (1 - a_0)^{m+1} - 1 + (1 - a_0)^m) = f_0 a_0 (1 - a_0)^m$$

ここに f_0, a_0 の推定量を代入した式を推定に使う

$$\hat{\Delta}(n+m) = \hat{f}_0 \hat{a}_0 (1 - \hat{a}_0)^{m+1}$$

$$\hat{a}_0 \hat{f}_0 = \frac{f_1}{n} \cdot \frac{(n-1)f_1}{(n-1)f_1 + 2f_2} = \frac{f_1}{n} (1 - \hat{a}_0) \quad \text{より}$$

$$\hat{\Delta}(n+m) = \frac{f_1}{n} (1 - \hat{a}_0)^{m+1} = \frac{f_1}{n} \left(1 - \frac{2f_2}{(n-1)f_1 + 2f_2}\right)^{m+1}$$

まとめて

$${}_q \hat{D} = \left\{ \sum_{k=0}^{n-1} \binom{q-1}{k} (-1)^k \sum_{i=1}^S \frac{x_i}{n} \frac{\binom{n-x_i}{k}}{\binom{n-1}{k}} + \sum_{m=0}^{\infty} \binom{q-1}{n+m} (-1)^{n+m} \frac{f_1}{n} \left(1 - \frac{2f_2}{(n-1)f_1 + 2f_2}\right)^{m+1} \right\}^{\frac{1}{1-q}}$$

$q = 0$ (種数) のとき

$${}^0\hat{D} = \left\{ \sum_{k=0}^{n-1} \binom{-1}{k} (-1)^k \hat{\Delta}(k) + \sum_{m=0}^{\infty} \binom{-1}{n+m} (-1)^{n+m} \hat{\Delta}(n+m) \right\}$$

$$\binom{-1}{k} = \frac{(-1)(-2)\cdots(-k)}{k!} = (-1)^k \quad \text{だから} \quad {}^0\hat{D} = \sum_{k=0}^{n-1} (-1)^k \hat{\Delta}(k) + \sum_{m=0}^{\infty} \hat{\Delta}(n+m)$$

次の等式が成り立ちます (証明は数学としてテクニカルなので後回し(たぶん省略))

$$\sum_{k=0}^{n-1} (-1)^k \hat{\Delta}(k) = S_{obs}$$

$$\hat{\Delta}(k) = \sum_{i=1}^S \frac{x_i}{n} \frac{\binom{n-x_i}{k}}{\binom{n-1}{k}}$$

第2項の無限和は、等比級数の和の公式を用いて

$$\sum_{m=0}^{\infty} \hat{\Delta}(n+m) = \frac{f_1}{n} \sum_{m=0}^{\infty} (1 - \hat{a}_0)^{m+1} = \frac{f_1}{n} \frac{1 - \hat{a}_0}{\hat{a}_0} = \frac{f_1}{n} \frac{1 - \frac{2f_2}{(n-1)f_1 + 2f_2}}{\frac{2f_2}{(n-1)f_1 + 2f_2}} = \frac{f_1}{n} \frac{(n-1)f_1}{2f_2} = \frac{(n-1)}{n} \frac{f_1^2}{2f_2}$$

結局、

$${}^0\hat{D} = S_{obs} + \hat{f}_0 = S_{obs} + \frac{n-1}{n} \frac{f_1^2}{2f_2}$$

観察種数に、観察されなかった種数についてのChaoの推定量を加えるだけでした

$q = 1$ Shannon

(b) Letting q tend to 1, we have

$$^1\hat{D} = \exp\left(\sum_{1 \leq X_i \leq n-1} \frac{X_i}{n} \left(\sum_{k=X_i}^{n-1} \frac{1}{k}\right) + \frac{f_1}{n} (1-A)^{-n+1} \left[-\log A - \sum_{r=1}^{n-1} \frac{(1-A)^r}{r}\right]\right). \quad (S1.5)$$

(c) For $q = 2, 3, \dots, n$, we have

$$^q\hat{D}^{1-q} = \sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^k \hat{\Delta}(k) = \sum_{X_i \geq q} \frac{X_i(X_i-1)\dots(X_i-q+1)}{n(n-1)\dots(n-q+1)}.$$

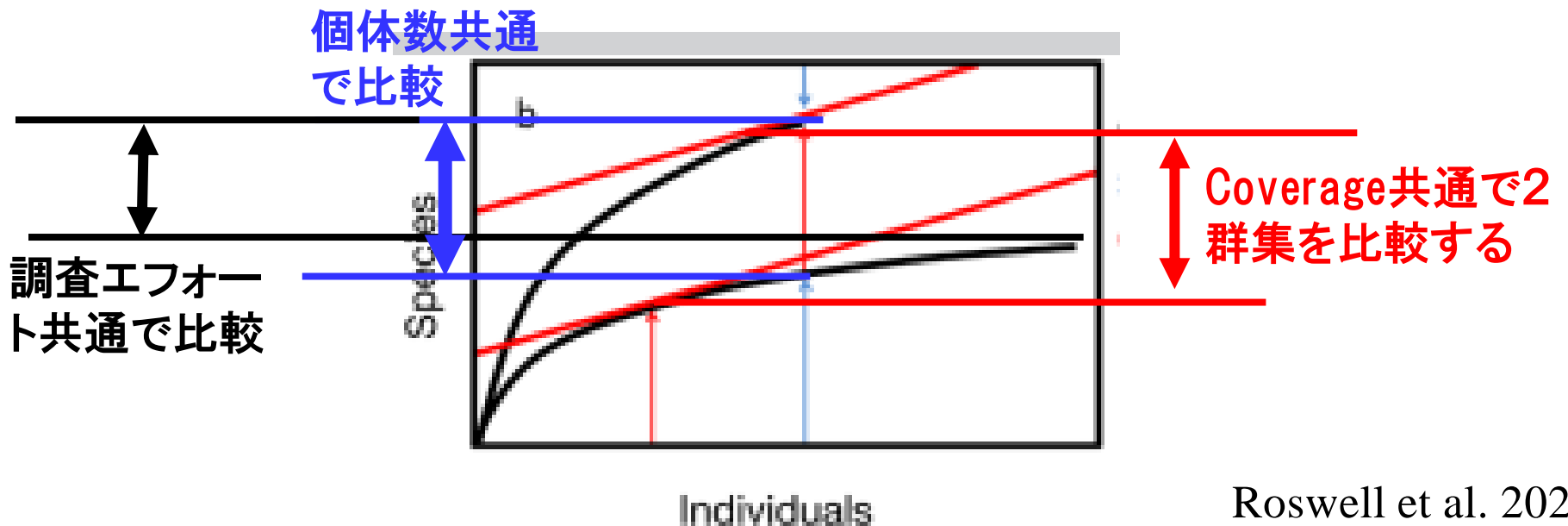
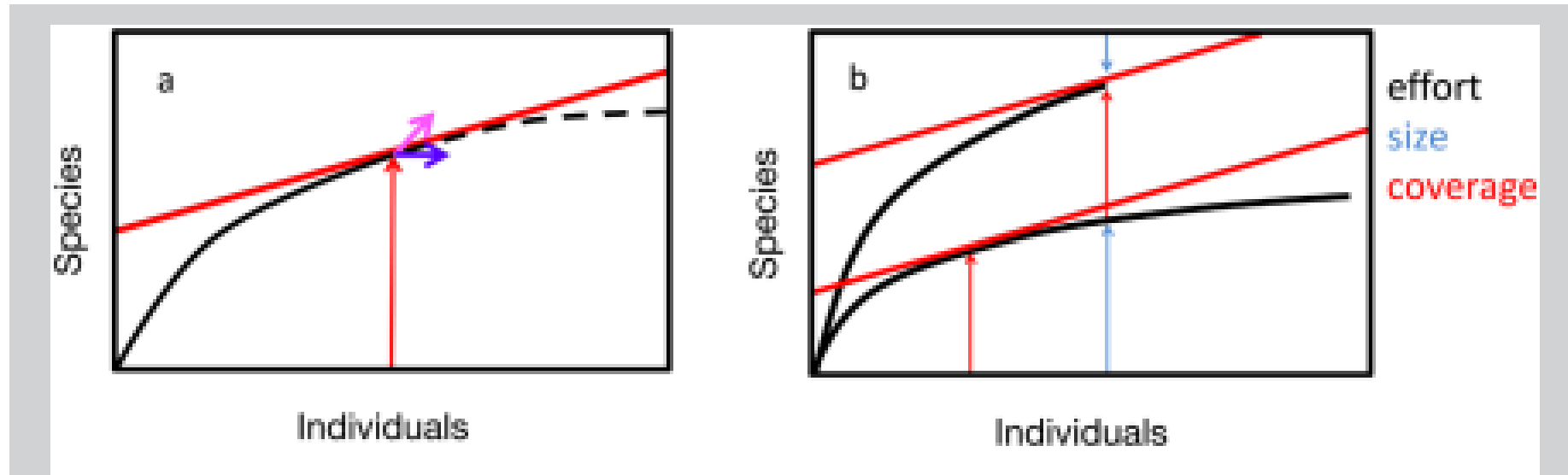
The right hand side of the above formula is positive only when $q \leq \max X_i$ under which we have the estimator

$$^q\hat{D} = \left(\sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^k \hat{\Delta}(k)\right)^{1/(1-q)} = \left(\sum_{X_i \geq q} \frac{X_i(X_i-1)\dots(X_i-q+1)}{n(n-1)\dots(n-q+1)}\right)^{1/(1-q)}. \quad (S1.6)$$

(d) For any q between 0 and $\max X_i$, the general formula for our estimator can be expressed as the following formula with finite terms:

$$^q\hat{D} = \left(\sum_{k=0}^{n-1} \binom{q-1}{k} (-1)^k \hat{\Delta}(k) + \frac{f_1}{n} (1-A)^{-n+1} \left[A^{q-1} - \sum_{r=0}^{n-1} \binom{q-1}{r} (A-1)^r\right]\right)^{1/(1-q)}. \quad (S1.7)$$

サンプル被覆度の応用例: 2つの群集の比較の基準



観察されなかった種数に関する Chao の推定量

.....

k 個体観察できた種の数: f_k

仮定から関係式を作り

.....

4個体観察できた種の数: f_4
3個体観察できた種の数: f_3
2個体観察できた種の数: f_2
1個体観察できた種の数: f_1
0個体観察できた種の数 = 観察できなかった種数: f_0

$$\mathbf{E}[F_0] \geq \frac{\left(\frac{\mathbf{E}[F_1]}{n}\right)^2}{\frac{2\mathbf{E}[F_2]}{n(n-1)}} = \frac{n-1}{n} \cdot \frac{(\mathbf{E}[F_1])^2}{2\mathbf{E}[F_2]}$$

観察値(データ)を入れる

$$\hat{f}_0 = \frac{n-1}{n} \cdot \frac{f_1^2}{2f_2}$$

ほかにも観察されなかった種数に関する推定量
がいろいろ提唱されています。

類似の問題: n 個体のサンプルをとったとき、

- ・ 次の1個体が新種である確率
- ・ 新たな m 個体の中の新種の数

方法はいくつかに大別できます

- 個体数—種数—個体数曲線 を何らかの数式で fitting させ、個体数 $\rightarrow \infty$ のときの極限
 どのような数式を fitting させるか？

- k 個体観察できた種の数 (f_k) を使って f_0 を推定する。Chaoの推定量は分母があるため推定値が荒れる。

線形式で表せないか？

- ベイズ推定

In the early 1940s, naturalist Corbet had spent 2 y trapping butterflies in Malaya. At the end of that time, he constructed a table (see below) to show how many times he had trapped various butterfly species. For example, 118 species were so rare that Corbet had trapped only one specimen of each, 74 species had been trapped twice each, etc.

Frequency	1	2	3	4	5	...	14	15
Species	118	74	44	24	29	...	12	6

Corbet returned to England with his table, and asked R. A. Fisher, the greatest of all statisticians, how many new species he would see if he returned to Malaya for another 2 y of trapping. This question seems impossible to answer, because it refers to a column of Corbet's table that doesn't exist, the "0" column. Fisher provided an interesting answer that was later improved on [by Good and Toulmin (17)]. The number of new species you can expect to see in 2 y of additional trapping is

$$118 - 74 + 44 - 24 + \dots - 12 + 6 = 75.$$

k 個体観察できた種の数 (f_k) を使って 観察されなかった種数 f_0 を推定する発想はフィッシャーに由来する

THE NUMBER OF NEW SPECIES, AND THE INCREASE IN POPULATION COVERAGE, WHEN A SAMPLE IS INCREASED

BY I. J. GOOD AND G. H. TOULMIN

A sample of size N is drawn at random from a population of animals of various species. Methods are given for estimating, knowing only the contents of this sample, the number of species which will be represented r times in a second sample of size λN ; these also enable us to estimate the number of different species and the proportion of the whole population represented in the second sample. A formula is found for the variance of the estimate; when $\lambda > 2$, this variance becomes in general very large, so that the estimate is useless without some modification. This difficulty can be partly overcome, at least for $\lambda < 5$, by using Euler's method with a suitable parameter or the methods described by Shanks (1955) to hasten the convergence of the series by which the estimate is expressed. The methods are applied to samples of words from *Our Mutual Friend*, to an entomological sample, and to a sample of nouns from Macaulay's essay on Bacon.

1. INTRODUCTION

We present here a further development of the theory expounded by Good (1953); that paper will be referred to, for brevity, by the letter G throughout.

We imagine a random sample of size N , the *basic sample*, to be drawn from an infinite population of animals of various species, and suppose that n_r distinct species are each represented exactly r times in the sample, so that

$$\sum_{r=1}^{\infty} r n_r = N. \quad (1)$$

We write

$$d = \sum_{r=1}^{\infty} n_r,$$

the total number of distinct species in the sample. It is convenient (though, as was pointed out in G, not essential) to suppose that the total number of distinct species in the population is a known finite number s , so that we can calculate

$$n_0 = s - d, \quad (2)$$

第2次大戦中にTuringが暗号解読のために考案した手法は観察されなかった種数の推定をイメージしていた。それを数学として定式化した論文のひとつ

Good-Toulmin記法: 黒
島谷記法: 青

$$n_r \Leftrightarrow f_k$$

N : サンプル数 n

d : 観察された種数

s : 真の種数 S

n_0 : 観察されなかった種数 f_0

Good and
Toulmin 1956

n サンプルを含めて計 tn 個体 ($m = (t-1)n$ 個のサンプルを追加する)

N λN

2. ESTIMATION OF $\mathcal{E}(n_r(\lambda))$

$F_k(t)$: 計 tn 個体の中で k 個体観察
される種数(確率変数)
 $E[F_k(t)]$: その期待値

Let $p_\mu (\mu = 1, 2, \dots, s)$ be the population frequencies of the s

p_s : 種 s の相対頻度

$$\mathcal{E}(n_r) = \sum_{\mu=1}^s \binom{N}{r} p_\mu^r (1 - p_\mu)^{N-r}$$

$$E[F_k] = \sum_{s=1}^S \binom{n}{k} p_s^k (1 - p_s)^{n-k}$$

(In G, the left-hand side is written as $\mathcal{E}_N(n_r|H)$. As explained above, we use the symbol n_r only with reference to the basic sample, of size N ; we omit the H , which refers to the hypothesis that the population frequencies are $\{p_\mu\}$, because we shall not be concerned with expectations on any other hypothesis.) For the second sample, we have similarly, assuming $p_\mu < \frac{1}{2}$ for all μ ,

$$\mathcal{E}(n_r(\lambda)) = \sum_{\mu=1}^s \binom{\lambda N}{r} p_\mu^r (1 - p_\mu)^{\lambda N-r}$$

$$E[F_k(t)] = \sum_{s=1}^S \binom{tn}{k} p_s^k (1 - p_s)^{tn-k}$$

$$= \sum_{\mu=1}^s \binom{\lambda N}{r} p_\mu^r (1 - p_\mu)^{N-r} \left(1 + \frac{p_\mu}{1 - p_\mu}\right)^{-(\lambda-1)N}$$

$$(x + y)^m = \sum_{k=0}^{\infty} \binom{m}{k} x^k y^{m-k}$$

$$= \sum_{\mu=1}^s \binom{\lambda N}{r} p_\mu^r (1 - p_\mu)^{N-r} \sum_{i=0}^{\infty} \binom{-(\lambda-1)N}{i} p_\mu^i (1 - p_\mu)^{-i}$$

(13)

$$= \sum_{i=0}^{\infty} \binom{\lambda N}{r} \binom{-(\lambda-1)N}{i} \sum_{\mu=1}^s p_\mu^{r+i} (1 - p_\mu)^{N-(r+i)}$$

$$= \sum_{i=0}^{\infty} \frac{\binom{\lambda N}{r} \binom{-(\lambda-1)N}{i}}{\binom{N}{r+i}} \mathcal{E}(n_{r+i})$$

$$E[F_{k+i}] = \sum_{s=1}^S \binom{n}{k+i} p_s^k (1 - p_s)^{n-(k+i)}$$

write

$$\frac{\binom{\lambda N}{r} \binom{-(\lambda-1)N}{i}}{\binom{N}{r+i}} = \frac{(\lambda N)^r (-(\lambda-1)N)^i (r+i)!}{r! i! N^{r+i}} = (-1)^i \lambda^r (\lambda-1)^i \binom{r+i}{r}. \quad (15)$$

高校数学

Hence

$$\mathcal{G}(n_r(\lambda)) \simeq \lambda^r \sum_{i=0}^{\infty} (-1)^i \binom{r+i}{r} (\lambda-1)^i \mathcal{G}(n_{r+i}), \quad (16)$$

the partial sums erring alternately in excess and defect.

(16)式を島谷記法で書くと

$$\mathbf{E}[F_k(t)] \approx t^k \sum_{i=0}^{\infty} (-1)^i \binom{k+i}{k} (t-1)^i \mathbf{E}[F_{k+i}]$$

左辺は tn 個のサンプルについての $F_k(t)$ の期待値, 右辺は n 個のサンプルのとき。
つまり、 tn サンプルと n サンプルの関係式が得られた。

n サンプルの時点でのデータから tn サンプルの時を推定できそう?

右辺の期待値 $\mathbf{E}[F_{k+i}]$ に n 個のサンプルのときの観察値 f_{k+i} を代入して、
左辺(tn 個のサンプル)の推定量にする(推定量では $\hat{}$ を付ける習慣があります)。
特に関心の高い $k=0$ のときだけ書くと

$$\hat{F}_0(t) = \sum_{i=0}^{\infty} (-1)^i (t-1)^i f_i = f_0 + \sum_{i=1}^{\infty} (-1)^i (t-1)^i f_i$$

右辺の中の f_0 は不明だから推定量になっていない???

$$S_{obs} - \sum_{i=1}^{\infty} (-1)^i (t-1)^i F_i$$

という確率変数(n サンプルのデータの実現値を代入すれば計算できる)の期待値について、

観察された種数 +
観察されなかった
種数 = 真の種数

$$\mathbf{E}[S_{obs} - \sum_{i=1}^{\infty} (-1)^i (t-1)^i F_i] = \mathbf{E}[S_{obs} + F_0 - F_0 - \sum_{i=1}^{\infty} (-1)^i (t-1)^i F_i]$$

$$\mathbf{E}[S_{obs} + F_0] = S \quad \text{だから}$$

$$= S - \mathbf{E}[\sum_{i=0}^{\infty} (-1)^i (t-1)^i F_i] \approx S - \hat{F}_0(t) \approx tn \quad \text{個体で観察される種数の推定量}$$

左辺に n サンプルのデータの実現値を代入

$$s_{obs} - \sum_{i=1}^{\infty} (-1)^i (t-1)^i f_i$$

n 個体では観察されず tn 個体に増やして初めて観察される種数は

$$s_{obs} - \sum_{i=1}^{\infty} (-1)^i (t-1)^i f_i - s_{obs} = \sum_{i=1}^{\infty} (-1)^{i+1} (t-1)^i f_i$$

$t = 2$ のときはまさしくFisherの式

$t > 1$ (新たなサンプル数がそれまでのより多い)のとき、右辺は発散、あるいは f_k のわずかな変化で大きく変動してしまう。

そこで、交代級数(係数が+と-を交互にとる)に関するオイラー変換を使って、収束する式に直す。収束をいかに早くするか、様々な論文が出ている。

Optimal prediction of the number of unseen species

Alon Orlitsky^a, Ananda Theertha Suresh^{b,1}, and Yihong Wu^c

^aElectrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093; ^bGoogle Research, New York, NY 10011; and ^cDepartment of Statistics, Yale University, New Haven, CT 06511

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved September 16, 2016 (received for review May 14, 2016)

Estimating the number of unseen species is an important problem in many scientific endeavors. Its most popular formulation, introduced by Fisher et al. [Fisher RA, Corbet AS, Williams CB (1943) *J Animal Ecol* 12(1):42–58], uses n samples to predict the number U of hitherto unseen species that would be observed if $t \cdot n$ new samples were collected. Of considerable interest is the largest ratio t between the number of new and existing samples for which U can be accurately predicted. In seminal works, Good and Toulmin [Good I, Toulmin G (1956) *Biometrika* 43(102):45–63] constructed an intriguing estimator that predicts U for all $t \leq 1$. Subsequently, Efron and Thisted [Efron B, Thisted R (1976) *Biometrika* 63(3):435–447] proposed a modification that empirically predicts U even for some $t > 1$, but without provable guarantees. We derive a class of estimators that provably predict U all of the way up to $t \propto \log n$. We also show that this range is the best possible and that the estimator's mean-square error is near optimal for any t . Our approach yields a provable guarantee for the Efron–Thisted estimator and, in addition, a variant with stronger theoretical and experimental performance than existing methodologies on a variety of synthetic and real datasets. The estimators are simple, linear, computationally efficient, and scalable to massive datasets. Their performance guarantees hold uniformly for all distributions, and apply to all four standard sampling models commonly used across various scientific disciplines: multinomial, Poisson, hypergeometric, and Bernoulli product.

In the early 1940s, naturalist Corbet had spent 2 y trapping butterflies in Malaya. At the end of that time, he constructed a table (see below) to show how many times he had trapped various butterfly species. For example, 118 species were so rare that Corbet had trapped only one specimen of each, 74 species had been trapped twice each, etc.

Frequency	1	2	3	4	5	...	14	15
Species	118	74	44	24	29	...	12	6

Corbet returned to England with his table, and asked R. A. Fisher, the greatest of all statisticians, how many new species he would see if he returned to Malaya for another 2 y of trapping. This question seems impossible to answer, because it refers to a column of Corbet's table that doesn't exist, the "0" column. Fisher provided an interesting answer that was later improved on [by Good and Toulmin (17)]. The number of new species you can expect to see in 2 y of additional trapping is

$$118 - 74 + 44 - 24 + \cdots - 12 + 6 = 75.$$

To predict U for $t > 1$, Good and Toulmin (17) suggested using the Euler transform (20) that converts an alternating series into another series with the same sum, and heuristically often converges faster. Interestingly, Efron and Thisted (5) showed that, when the Euler transform of U^{GT} is truncated after k terms, it can be expressed as another simple linear estimator,

$$U^{\text{ET}} \triangleq \sum_{i=1}^n h_i^{\text{ET}} \cdot \underline{\Phi_i}, \quad \begin{array}{l} \text{\textcolor{red}{k個体観察できた種の数 } } f_k \\ \text{\textcolor{red}{[2]}} \end{array}$$

where

$$\hat{n}_0(t) = \sum_{i=0}^{\infty} (-1)^i (t-1)^i f_i = n_0 + \sum_{i=1}^{\infty} (-1)^i (t-1)^i f_i$$

$$h_i^{\text{ET}} \triangleq -(-t)^i \cdot \mathbb{P}\left(\text{Bin}\left(k, \frac{1}{1+t}\right) \geq i\right),$$

$m = tn$ 個のサンプルを追加

and

Bin = binomial 2項分布

$$\mathbb{P}\left(\text{Bin}\left(k, \frac{1}{1+t}\right) \geq i\right) = \begin{cases} \sum_{j=i}^k \binom{k}{j} \frac{t^{k-j}}{(1+t)^k} & i \leq k, \\ 0 & i > k, \end{cases}$$

is the binomial tail probability that decays with i , thereby moderating the rapid growth of $(-t)^i$.

Small-Sample Estimation of Species Richness Applied to Forest Communities

Wen-Han Hwang* and Tsung-Jen Shen**

Department of Applied Mathematics and Institute of Statistics,
National Chung Hsing University, Taichung, Taiwan

**email:* wenhan@nchu.edu.tw

***email:* tjshen@nchu.edu.tw

SUMMARY. Many well-known methods are available for estimating the number of species in a forest community. However, most existing methods result in considerable negative bias in applications, where field surveys typically represent only a small fraction of sampled communities. This article develops a new method based on sampling with replacement to estimate species richness via the generalized jackknife procedure. The proposed estimator yields small bias and reasonably accurate interval estimation even with small samples. The performance of the proposed estimator is compared with several typical estimators via simulation study using two complete census datasets from Panama and Malaysia.

KEY WORDS: Abundance; Biodiversity; Generalized jackknife; Quadrat sampling; Species richness.

1. Introduction

In many ecological field studies, it is essential to know the number of species in a community. This so-called species richness is also the simplest, most fundamental, and intuitive concept of biodiversity. Applying to various disciplines, species richness can be the population of taxicabs in a city (Carothers, 1973), the total number of words Shakespeare knew (Efron and Thisted, 1976), the population size for a specific animal species (Burnham and Overton, 1978), the number of bugs in a software program (Chao, Ma, and Yang, 1993), the

data, respectively. The Chao1 and Chao2 estimators are derived as lower bound estimators for species richness, however, though they work well and serve as species richness estimators in many empirical studies (see, for example, Chazdon et al., 1998; Gimaret-Carpentier, Chessel, and Pascal, 1998). Subsequently, Chao and Lee (1992) and Chao, Lee, and Jeng (1992) used the idea of sample coverage and proposed abundance- and incidence-based coverage estimators (ACE and ICE), respectively. More recently, Norris and Pollock (1998); Colwell, Mao, and Chang (2004); and Mao and Lindsay (2007)

ChaoやGood-Turlmin
同様、 k 個体観察でき
た種の数 (f_k) を使っ
て 観察されなかった
種数 f_0 を推定する別
な方法のひとつ

Chao1 estimator (Chao, 1984). Instead of providing a surrogate of K directly, a lower bound estimate for K is considered here. By the Cauchy–Schwarz inequality, we have

$$K = \frac{\sum_{i=1}^S (1 - p_i)^n}{\sum_{i=1}^S p_i (1 - p_i)^{n-1}} \geq \frac{\sum_{i=1}^S p_i (1 - p_i)^{n-1}}{\sum_{i=1}^S p_i^2 (1 - p_i)^{n-2}} = \frac{(n-1)E(f_1)}{2E(f_2)}. \quad (6)$$

This implies a lower bound of species richness through equation (2), i.e., $S \geq E(D_n) + (n-1)\{E(f_1)\}^2/\{2nE(f_2)\}$. Using the method of moments and ignoring the term $(n-1)/n$, we obtain the lower bound estimate

$$\hat{S} = D_n + \frac{f_1^2}{2f_2},$$

which was first proposed in Chao (1984). Since this formula is undefined when $f_2 = 0$, we recommend using the bias-corrected version, $\hat{S} = D_n + f_1(f_1 - 1)/(2(f_2 + 1))$, instead. In addition, it is interesting to note that, though this estimator was intended to yield a lower bound of S , the bound is really sharp (Chao, 2005) and hence could serve as an estimate of S in practical applications.

Proposed estimator. We follow a philosophy similar to that in the derivation of the Chao1 estimator. The use of the Cauchy–Schwarz inequality in (6) could be generalized to

$$K = \frac{nE(f_0)}{E(f_1)} \geq \frac{(n-1)E(f_1)}{2E(f_2)} \geq \frac{(n-2)E(f_2)}{3E(f_3)} \geq \dots \quad (7)$$

Thus, we are convinced that the regression function should be decreasing with an attenuating tangent slope. We therefore propose the exponential regression model

$$y_i = \beta_0 \exp(\beta_1 i^{\beta_2}) + \epsilon_i, \quad (8)$$

where $i = 1, \dots, n-1$, $\beta_0 > 0$, $\beta_1 < 0$, $\beta_2 > 0$, and ϵ_i denotes random errors. Since we have the observed y_i on hand, we can estimate β_0, β_1 , and β_2 using a standard nonlinear regression routine. It follows that $\hat{K} = \hat{\beta}_0$ and the induced estimator of species richness is

$$\hat{S} = D_n + \hat{\beta}_0 \frac{f_1}{n}.$$

Chaoの方法を拡張して f_1 ,
 f_2 , ...の不等式を作り、こ
れらを回帰することで観察
されなかった種数 f_0 を推
定する

方法はいくつかは大別できます

- ・個体数—種数—個体数曲線 を何らかの数式でfittingさせ、個体数 $\rightarrow \infty$ のときの極限
- ・ k 個体観察できた種の数 (f_k) を使って f_0 を推定する。Chaoの推定量は分母があるため推定値が荒れる。

・ベイズ推定

$$p_1 = V_1$$

$$p_s = (1 - V_1)(1 - V_2) \cdots (1 - V_{s-1})V_s$$

などで(無限個)の相対頻度に事前分布(ハイパーパラメータ含む)を与える。

- ・観察されなかった種数などの事後分布を計算する。
- ・データからハイパーパラメータを最適化する(経験ベイズ)。

Hill 数以外の多様性指数：データからうまく推定できる指数(不偏推定量が開発されている)

Generalized Simpson entropy $\zeta_r = \sum_{s=1}^S p_s (1 - p_s)^r \quad r = 1, 2, \dots$

$r + 1$ 番目の個体为新種である確率

Hurlbert diversity index ${}_k S = \sum_{s=1}^S (1 - (1 - p_s)^k) \quad k = 2, 3, \dots$

ランダムに選んだ k 個体の中の種数の期待値

不偏推定量(推定量の期待値＝真値)が知られている

$$\hat{\zeta}_r = \sum_{s=1}^{S'} \hat{p}_s \prod_{j=1}^r \left(1 - \frac{N_s - 1}{N - j}\right)$$

$${}_k \hat{S} = \sum_{s=1}^S \left(1 - \frac{(N - N_s)(N - N_s - 1) \cdots (N - N_s - k + 1)}{N(N - 1) \cdots (N - k + 1)}\right)$$

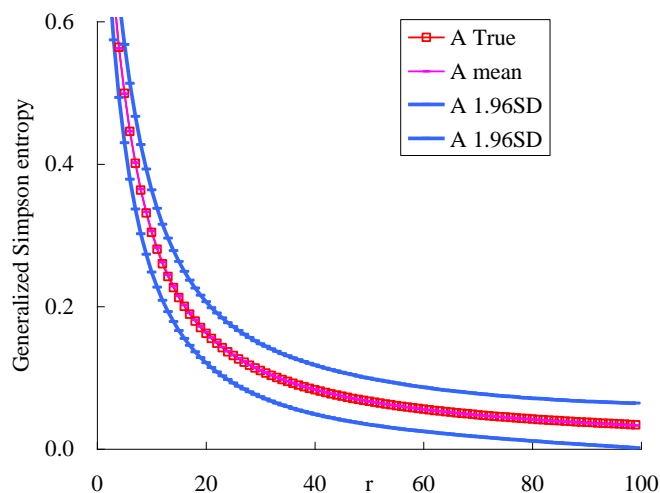
Generalized Simpson entropy

$$\zeta_r = \sum_{s=1}^S p_s (1 - p_s)^r$$

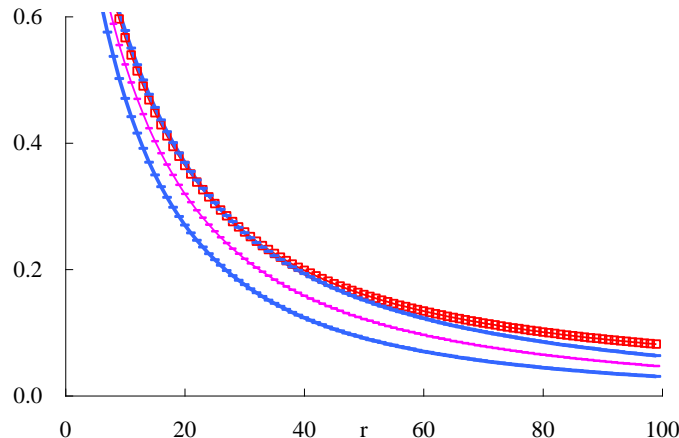
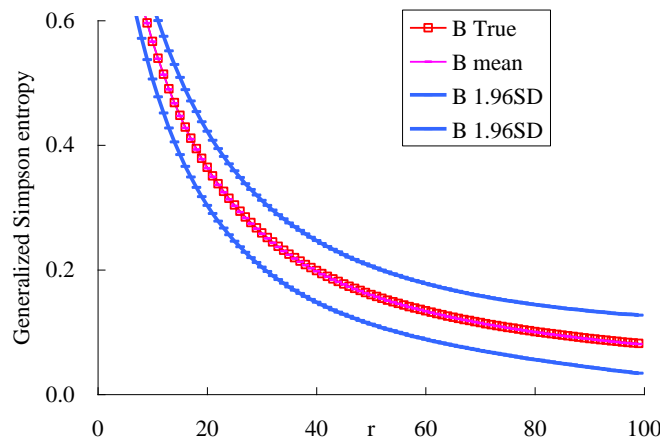
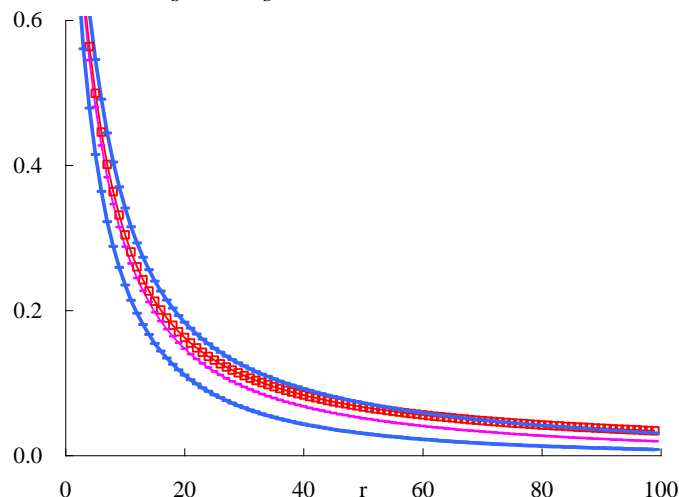
いわゆる割り算では過少推定だが(右)、不偏推定量を使えば確かに真値の上下に均等に散らばっている(左)

Examples of simulations (100 samples).

Unbiased estimator was used.



$\hat{p}_s = N_s / N$ was used.

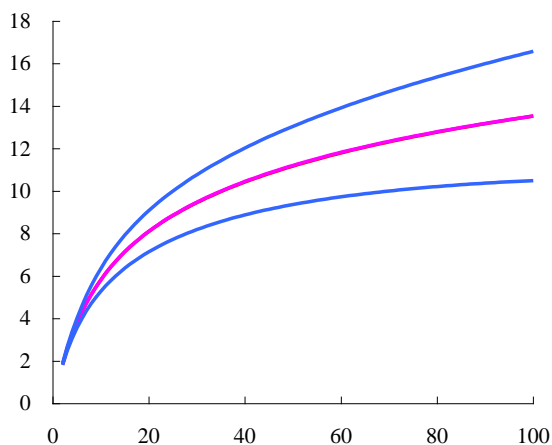


Hurlbert diversity ${}^k S = \sum_{s=1}^S (1 - (1 - p_s)^k)$

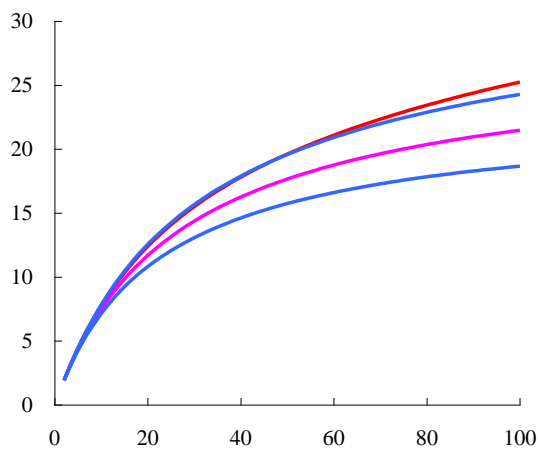
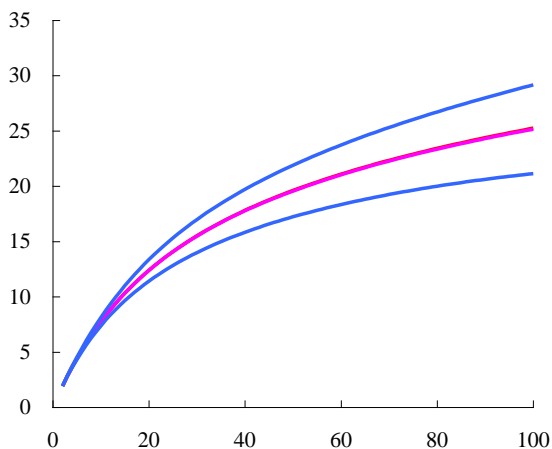
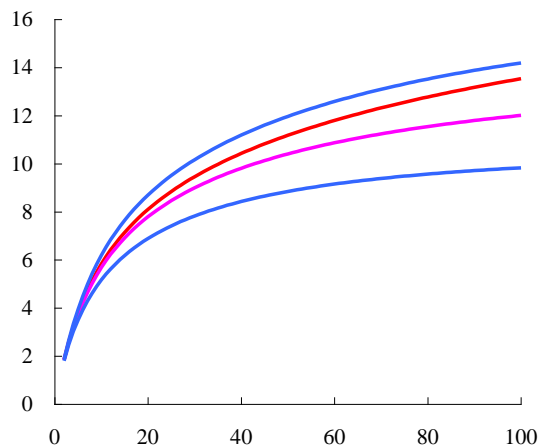
いわゆる割り算では過少推定だが(右)、不偏推定量を使えば確かに真値の上下に均等に散らばっている(左)

Examples of
simulations
(100 samples).

Unbiased estimator was used.



$\hat{p}_s = N_s / N$ was used.



Hurlbert 指数を「有効な種数」に変換 $p_s = 1/S$ のとき種数と等しくなるようにする

TABLE 1 Comparing Hill numbers, individual-based

	Hill numbers	Individual-based rarefaction	ENS rarefaction
Symbol	qD	S_n	E_n
Formula	${}^qD = \left(\sum_{i=1}^S p_i^q \right)^{\frac{1}{1-q}}$ Equation (1)	$S_n = S - \sum_{X_i \geq 1} \frac{\binom{N-X_i}{n}}{\binom{N}{n}}$ Equation (2)	$S_n = E_n \left(1 - \left(1 - \frac{1}{E_n} \right)^n \right)$ Equation (3)
Range	1, N	1, n	1, ∞
ENS	Yes	No	Yes
Estimation bias	Downward bias for $q < 2$	Unbiased	Unbiased
Description	ENS transformation ("true diversity") of any diversity index that is a function of $\sum_{i=1}^S p_i^q$ (e.g. Richness ($q = 0$), Shannon ($q = 1$), Simpson ($q = 2$)); Defined as the species richness of a hypothetical perfectly even community that has the same diversity index value as the sample	The expected species richness of a sample of n individuals ($n < N$)	ENS transformation of S_n . Defined as the species richness of a hypothetical community that has the same rarefied richness (S_n) as the sample and infinitely many individuals
Influence of relative abundances	The higher q , the lower the influence of rare species	The higher n , the higher the influence of rare species	The higher n , the higher the influence of rare species
References	Hill (1973), Jost (2006)	Hurlbert (1971), Gotelli and Colwell (2001)	Dauby and Hardy (2012)

Note: S , observed species richness; p_i , relative abundance of species i ; q , exponent that determines the sensitivity to rare species (0 = very sensitive, 2 = not very sensitive); N , observed number of individuals in the sample; X_i , number of individuals of species i ; ENS, the effective number of species which is the number of equally abundant species that results in the same value of diversity as the sample. To calculate E_n , Equation 3 can be solved numerically for given values of S_n and n .

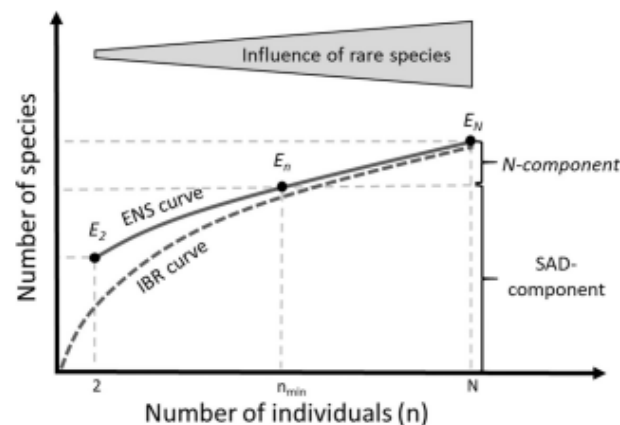


FIGURE 1 Schematic drawing of an individual-based rarefaction (IBR) curve and the corresponding effective number of species (ENS) curve. The IBR curve is constrained by the values of n (i.e. it is bound to start at the $x = y = 1$), when the ENS curve is unconstrained on the vertical axis. The ENS for a standardized number of individuals E_n reflects the "component" in our framework. The difference between the diversity (ENS_N) and the SAD-component (ENS_n) results in the fact that samples usually exceed the number of individuals used for standardization. As this portion of the total diversity change reflects abundance variation, we call it "N-comp

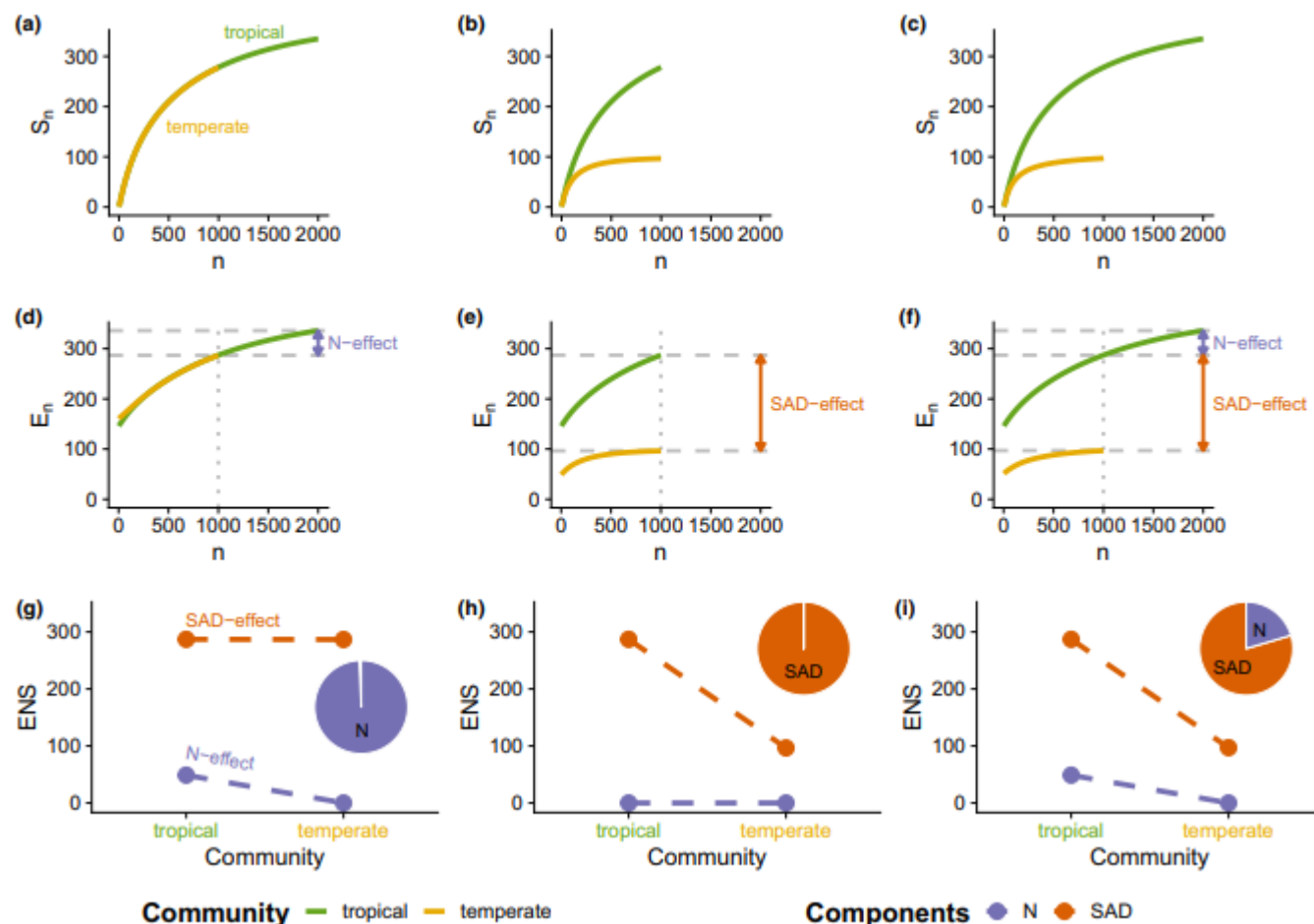
How does variation in total and relative abundance contribute to gradients of species diversity?

Thore Engel^{1,2} | Shane A. Blowes^{1,2} | Daniel J. McGlinn³ | Nicholas J. Gotelli⁴ | Brian J. McGill⁵ | Jonathan M. Chase^{1,2}

More individuals

SAD change

Both








REVIEW

Ecological Monographs, 91(3), 2021, e01452

© 2021 The Authors. *Ecological Monographs* published by Wiley Periodicals LLC on behalf of Ecological Society of America.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Measurement and analysis of interspecific spatial associations as a facet of biodiversity

PETR KEIL ^{1,2,3,7} THORSTEN WIEGAND ^{1,4} ANIKÓ B. TÓTH ⁵
DANIEL J. MCGLINN ⁶ AND JONATHAN M. CHASE ^{1,2}

¹German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

²Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany

³Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Praha – Suchbátka, 165 00, Czech Republic

⁴Department of Ecological Modelling, Helmholtz Centre for Environmental Research - UFZ, 04318 Leipzig, Germany

⁵Centre for Ecosystem Sciences, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052 Australia

⁶Department of Biology, College of Charleston, Charleston, South Carolina 29401 USA

Citation: Keil, P., T. Wiegand, A. B. Tóth, D. J. McGlinn, and J. M. Chase. 2021. Measurement and analysis of interspecific spatial associations as a facet of biodiversity. *Ecological Monographs* 91(3):e01452. 10.1002/ecm.1452

Abstract. Interspecific spatial associations (ISA), which include co-occurrences, segregations, or attractions among two or more species, can provide important insights into the spatial structuring of communities. However, ISA has primarily been examined in the context of understanding interspecific interactions, while other aspects of ISA, including its relations to other biodiversity facets and how it changes in the face of anthropogenic pressures, have been largely neglected. This is likely because it is unclear what makes ISA useful

In contrast to the spatially implicit indices for binary and abundance data, the connection between ISA and beta diversity is well known in analyses of point patterns (Wiegand and Moloney 2014). The ISA-beta connection can be demonstrated in the spatially-explicit version of Simpson's evenness index $\beta(r)$ (Shimatani 2001, Wiegand and Moloney 2014: section 3.1.5.1). Unlike the traditional spatially implicit version of the Simpson's index (Simpson 1949; i.e., the probability that two randomly selected individuals are heterospecifics), which is a measure of evenness, $\beta(r)$ is a measure of beta diversity, since it captures dissimilarity over a given distance (Shimatani 2001; i.e., the probability that two randomly selected individuals distance r apart are heterospecifics). The index is defined as

$$\beta(r) = \sum_{i=1}^S \sum_{\substack{j=1 \\ j \neq i}}^S f_i f_j \frac{g_{ij}(r)}{g(r)} = 1 - \sum_{m=1}^S \frac{f_m^2 g_{mm}(r)}{g(r)}. \quad (2)$$

We thus conclude that point pattern analysis, through $\beta(r)$, offers a comprehensive framework that can link abundances, CSA, ISA, gamma diversity, and alpha diversity, each with an exactly defined and mathematically tractable metrics. Not only does it stress the importance of making all of the diversity facets spatially explicit, but it also potentially offers a roadmap for future unification macroecology that deals with spatially implicit data on abundances or incidences.

ISA, species–area relationships, and species–accumulation curves

左は空間点過程によるアルファ多様性と β 多様性について数学の概念として定義付けようという試みです。

5年も前からこの発展版を考えていて、成功したら来年にでも公表します

species–area relationships (SAR), which are exactly related the Whittaker index over a continuous range of

右は種数–面積曲線の空間点過程版です。

When every individual's spatial position and identity is known, point pattern analysis also makes it clear that there is no direct link between ISA and SAR. The relevant equation is (Shimatani and Kubota 2004)

$$S(r) = \sum_{i=1}^Y H_i(r) \quad (3)$$

where $S(r)$ is number of species present within r from an arbitrarily chosen “test” location, H_i is the spherical contact distribution function for species i , which is the probability that the first neighbor of species i is distance r away from the test location. $S(r)$ becomes a species–area curve when r is converted to πr^2 . Importantly, the H_i is insensitive to ISA, since it is only based on the locations of species i . We note that point pattern analysis also has a scaling curve that is sensitive to ISA: the individual species–area relationship (ISAR; Wiegand et al. 2007),