帯大研報Ⅱ，8－1（1990）：1～7

α－エントロピーを用いた多様性推定について

種市信裕
（教養課程統計学教室）

1989年10月31日受理

On the estimation of diversity using α－entropy

Nobuhiro TANEICHI

摘　要

従来多様性の指数として用いられてきたShannon－エントロピー関数やGini-index を，それぞれ$\alpha=1$，$\alpha=2$の場合として含むHavrda-Charvat のα－エントロピー関数を多様性の指数として用いる場合における，多様性の指数を推定するうえでのバイアス修正，及び推定量の分散に基づく指数選択の基準について考察する。

1. 多様性の指数

m個の種とそれに属する個体数N_1，……，N_mが与えられた時，種の数 m（richness）と，それぞれの種に属する個体数の均等性（evenness）を同時に測る尺度として，多様性の指数は，生物学，特に生態学，遺伝学等でよく用いられている。多様性の指数として有名な尺度としては，Shannon－エントロピー関数 $H_{k,1}$ 及び，Gini－index $H_{k,2}$ がある。Shannon－エントロピー関数 $H_{k,1}$ は，

$$S_k=\{p=(p_1,\cdots\cdots,p_k)\mid \sum_{i=1}^{k}p_i=1,\ p_i\geqq 0\ (i=1,\cdots,k)\}$$

上で定義される関数として，

$$H_{k,1}(p) = -\sum_{i=1}^{k} p_i \log_e p_i$$

と表される。Gini−index $H_{k,2}$ は，S_k 上で定義される関数として，

$$H_{k,2}(p) = 1 - \sum_{i=1}^{k} p_i^2$$

と表される。$H_{1,k}$，$H_{2,k}$ には，次のような性質がある。

1. $H_k(p) = 0 \Leftrightarrow \exists i \ p_i = 1$
2. $\forall p \in S_k$ に対して $H_k(p) \leqq H_k(1/k, 1/k, \cdots\cdots, 1/k)$
3. $H_k(1/k, \cdots\cdots, 1/k) \leqq H_{k+1}((1/(k+1), \cdots\cdots, 1/(k+1))$

1，2の性質により，これらの尺度は均等性（evenness）を，3の性質により，種の数（richness）を示す尺度として用いられる。ところが1, 2, 3の性質は，$\alpha = 1$ の場合としてShannon−エントロピー，$\alpha = 2$ の場合として，Gini−indexとなる α −エントロピー関数

$$H_{k,\alpha}(p) = \begin{cases} (\alpha - 1)^{-1}(1 - \sum_{i=1}^{k} p_i^{\alpha}) & (\alpha \neq 1, \ \alpha > 0) \\ -\sum_{i=1}^{k} p_i \log_e p_i & (\alpha = 1) \end{cases}$$

(Havrda and Charvat, 1967)

全体が満たす性質である。このことからオーダー α のエントロピー関数全体は，多様性の指数としての条件を満たしていることがわかる。Rao,C.R.（1982），Nayak, T.K.（1984）らによって，多様性の指数としてオーダー α のエントロピーの関数の研究が行なわれているが，どの α を用いることが好ましいかについては考察されていない。

m個の母集団があり，それぞれの母集団はk個のカテゴリーがあるとする。m個の母集団からのそれぞれの標本数を $N_1, \cdots\cdots, N_m$ として，N_i (i=1,…, m) 個の標本の中でj番目（$j = 1, \cdots\cdots, k$）のカテゴリーに反応している個体を n_{ij} とする。$X_i = (n_{i1}, \cdots\cdots\cdots, n_{ik})$，$(i = 1, \cdots\cdots\cdots, m)$ とおき，$X_i \sim$ Mult (N_i, p_i) ただし，$p_i = (p_{i1}, \cdots\cdots, p_{ik})$，とすると，それぞれの母

集団における多様性の指数の M.L.E. は，

$$\widehat{H_{k,\alpha}(p_i)} = (\alpha-1)^{-1}(1-\Sigma_{j=1}^{k}\hat{p}_{ij}^{\alpha}),\quad (i=1,\cdots\cdots,m)$$

ただし，$\hat{p}_{ij}=(n_{ij}/N_i)$ となる。実際の問題ではこの推定値を用いて，m 個の母集団の多様性の指数の値を比較している。

2. 推定量のバイアス修正

m個の母集団における推定量の期待値 $E(\widehat{H_{k,\alpha}(p_i)})$ は次のように展開される。

$$E(\widehat{H_{k,\alpha}(p_i)}) = H_{k,\alpha}(p_i) + (1/N_i)A_{1i}(p_i) + (1/N_i^2)A_{2i}(p_i) + (1/N_i^3)A_{3i}(p_i) + O(1/N_i^4),\quad (i=1,\cdots,m)$$

ただし，

$$A_{1i}(p_i) = (\alpha/2)\{\Sigma_{j=1}^{k} p_{ij}^{\alpha-1}(p_{ij}-1)\}$$

$$A_{2i}(p_i) = \{\alpha(\alpha-2)/24\}\{\Sigma_{j=i}^{k} p_{ij}^{\alpha-2}(p_{ij}-1)\times\{-5+3\alpha+p_{ij}(1-3\alpha)\}\}$$

$$A_{3i}(p_i) = \{\alpha(\alpha-2)(\alpha-3)/48\}[\Sigma_{j=1}^{k} p_{ij}^{\alpha-3}(p_{ij}-1)\times\{(\alpha-2)(\alpha-3)+4(\alpha-1)p_{ij}+\alpha(\alpha-1)p_{ij}^2\}]$$

$p_i=(p_{i1},\cdots\cdots,p_{ik})$ は，各母集団においてそれぞれカデゴリーに反応する真の確率。

ここで，$H_{k,\alpha}(p_i)$ の推定量を

$$\theta_{i\alpha} = H_{k,\alpha}(p_i) - (1/N_i)A_{1i}(p_i) - (1/N_i^2)A_{2i}(p_i) - (1/N_i^3)A_{3i}(p_i)$$

とおくと,

$$E(\theta_{i\alpha}) = H_{k,\alpha}(p_i) - O(1/N_i^4)$$

$\theta_{i\alpha}$ の近似として

$$\hat{\theta}_{i\alpha} = H_{k,\alpha}(\hat{p}_i) - (1/N_i) A_{1i}(\hat{p}_i) - (1/N_i^2) A_{2i}(\hat{p}_i) - (1/N_i^3) A_{3i}(\hat{p}_i)$$

とおき，これを $H_{k,\alpha}(p_i)$ の推定量とする。（この推定量を用いると少なくとも $1/N_i$ のオーダーまでバイアスは修正されている。）

3. 多様性の指数としてどのαを用いれば良いか決める基準

本報告ではαを決める基準として，多様性の指数を推定するときの推定量の分散を用いることを考える。$H_{k,\alpha}(p_i)$ の推定量 $\hat{\theta}_{i\alpha}$ の分散の近似として

$$Var(\theta_{i\alpha}) \approx Var(\theta_{i\alpha}) = E\{\theta_{i\alpha} - E(\theta_{i\alpha})\}^2$$

とおくと,

$$N_i E\{\theta_{i\alpha} - E(\theta_{i\alpha})\}^2 = B_{1i}(p_i) + (1/N_i) B_{2i}(p_i) + (1/N_i^2) B_{3i}(p_i) + O(1/N_i^3) \quad (i = 1, \cdots\cdots, m)$$

ただし,

$$B_{1i}(p_i) = \{\alpha/(\alpha-1)\}^2 \{\Sigma_{j=1}^{k} p_{ij}^{2\alpha-1} - (\Sigma_{j=1}^{k} p_{ij}^{\alpha})^2\}$$

$$B_{2i}(p_i) = \{\alpha^2/2(\alpha-1)\}\{3(\alpha-1)\Sigma_{j=1}^{k} p_{ij}^{2\alpha-2} + 2(1-2\alpha)\times \Sigma_{j=1}^{k} p_{ij}^{2\alpha-1} - 2(\alpha-1)\Sigma_{i=1}^{k}\Sigma_{l=1}^{k} p_{ij}^{\alpha} p_{il}^{\alpha-2} + (3\alpha-1)(\Sigma_{j=1}^{k} p_{ij}^{\alpha})^2\}$$

$$B_{3i}(p_i)=\{\alpha^3/24(\alpha-1)\}\{2(14\alpha^3-60\alpha^2+87\alpha-43)\Sigma_{j=1}^{k} p_{ij}^{2\alpha-3}$$
$$+\{-36(\alpha-1)^2(2\alpha-3)\}\Sigma_{j=1}^{k} p_{ij}^{2\alpha-2}$$
$$+4(12\alpha^3-32\alpha^2+25\alpha-6)\Sigma_{j=1}^{k} p_{ij}^{2\alpha-1}$$
$$+\{-2(\alpha-2)^2(3\alpha-5)\}\Sigma_{j=1}^{k}\Sigma_{l=1}^{k} p_{ij}^{\alpha} p_{il}^{\alpha-2}$$
$$+12(\alpha-1)^2(3\alpha-4)\Sigma_{j=1}^{k}\Sigma_{l=1}^{k} p_{ij}^{\alpha} p_{il}^{\alpha-1}$$
$$+4(-7\alpha^3+16\alpha^2-9\alpha+1)(\Sigma_{j=1}^{k} p_{ij}^{\alpha})^2$$
$$+\{-6(\alpha-1)^3\}(\Sigma_{j=1}^{k} p_{ij}^{\alpha-1})^2\},$$

$p_i=(p_{i1},\cdots\cdots p_{ik})$は，各母集団においてそれぞれのカテゴリーに反応する真の確率。

各母集団からの標本数が比較的大きいときにはB_{1i}，比較的小さいときにはB_{2i}，B_{3i}まで考慮にいれることにする。ここでエントロピーの推定量の大きさで$\mathrm{Var}(\theta_{i\alpha})$を標準化したものを，

$$C_{i\alpha}(p_i)=\mathrm{Var}(\theta_{i\alpha})/(\hat{\theta}_{i\alpha})^2,\quad(i=1,\cdots\cdots m)$$

とおき，

$$C_{i\alpha}(\hat{p}_i)=\mathrm{Var}(\theta_{i\alpha})/(\hat{\theta}_{i\alpha})^2 \mid p_i=\hat{p}_i$$

として，

$$\min_{\alpha}\{\max\{C_{1\alpha}(\hat{p}_1),\cdots\cdots,C_{m\alpha}(\hat{p}_m)\}\} \qquad (3.1)$$

をαの値を決める基準とする。つまり，各母集団（1～m）における多様性の指数の推定量の分散の近似を推定値の大きさで標準化し，その最大値が最小となるようにαの値を定めることとする。

4. 数 値 例

表1で与えられる，生態学における生活形スペクトル（緯度系列）を例にとり多様性の指数$H_{k,1}(p)$（Shannon－エントロピー），$H_{k,2}(p)$（Gini－index），$H_{k,3}(p)$を比較する。

表1. 生活形スペクトル（緯度系列）

	地上植物	地表植物	半地表植物	地中植物	一年生植物
亜熱帯林	0.65	0.17	0.02	0.05	0.10
温暖帯林	0.54	0.09	0.24	0.09	0.04
冷温帯林	0.10	0.17	0.54	0.12	0.07
ツンドラ	0.01	0.22	0.60	0.15	0.02

いま，標本数は$N_1=N_2=N_3=N_4=100$とした場合を考える，それぞれの地域の推定確率ベクトルは，

$p_1=(0.65,\ 0.17,\ 0.20,\ 0.05,\ 0.1)$，

$p_2=(0.54,\ 0.09,\ 0.24,\ 0.09,\ 0.04)$，

$p_3=(0.1,\ 0.17,\ 0.54,\ 0.12,\ 0.07)$，

$p_4=(0.01,\ 0.02,\ 0.6,\ 0.15,\ 0.02)$，

であり，一般に用いられているShannon－エントロピー関数$H_{5,1}$を多様性の指数とすると，バイアス修正後の指定値は，$\hat{\theta}_{11}=1.053$，$\hat{\theta}_{21}=1.2379$，$\hat{\theta}_{31}=1.3052$，$\hat{\theta}_{41}=1.0509$となり，多様性の指数の大小関係は$\hat{\theta}_{41}<\hat{\theta}_{11}<\hat{\theta}_{21}<\hat{\theta}_{31}$となる，つまり多様性の大小関係は，ツンドラ＜亜熱帯林＜温暖帯林＜冷温帯林，となる。一方，標準化した推定分散を用いたαを決定する基準（3.1）で$\alpha=1$，2，3の場合を比較すると，分散として漸近分散B_{1i}のみを考えた場合，$1/N_i^2$のオーダーまで考えた場合，$1/N_i^3$のオーダーまで考えた場合，いずれも$\alpha=3$の場合が好ましいという結果がでる。そこで，このH$_{k,3}$のバイアス修正後の推定値を求めると，$\hat{\theta}_{13}=0.3595$，$\hat{\theta}_{23}=0.4136$

$\hat{\theta}_{33} = 0.4172$, $\hat{\theta}_{43} = 0.3849$ となり，多様性の指数の大小関係は$\hat{\theta}_{13} < \hat{\theta}_{43} < \hat{\theta}_{23} < \hat{\theta}_{33}$であることにより，多様性の大小関係は，亜熱帯林＜ツンドラ＜温暖帯林＜冷温帯林となり，推定分散の基準の意味では，従来の Shannon－エントロピーによる多様性の関係よりこちらの関係のほうが適切と思われる。

〔参考文献〕

1) Havrda, J. and Charvat, F. (1967). Quantification method in classification processes;concept of α－entropy, Kybernetika, 33, 30－35.

2) Nayak, T. K. (1984). On diversity measure based on entropy functions, Tech. Rept. No 84－16. Univ. of Pittsburgh.

3) Pielou, E. C. (1969). An introduction to mathematical ecology, John Wiley and Sons.

4) Rao, C. R. (1982). Diversity;its measurement, decomposition, apportionment and analysis, Sankya A, 44, 1－22.