生物多様性指数の推定：観察されなかった種

適当に定めた生物集団の中に生息する種とその個体数の調査は世界各地で盛んに行われ、データが蓄積されている。通常、集団の全個体を調べ上げることはできないから、一部の個体あるいは一部の領域を調べ、観察された種とその個体数を記録する。集団全体の多様性は、このデータから推定することになる。

自明なことではあるが、種数という多様性種数は、観察された種数を用いる限り、常に過少推定となる。どんなに頑張って調査しても、観察されなかった種は出るだろうし、そもそも、観察できなかった種を我々は知る術がない。これはもう原理的にやむを得ないとあきらめてしまうのが人の常であろう。

ところで、統計学に求められる重要な役割のひとつが、限られたサンプルから全体（母集団）を推定することである。しかし、データがない状況で、推定することは原理的に不可能である。観察されなかった種についての情報はないので、それらのリストを作る（推定する）ことは不可能である。

ところが、観察されなかった種名（種のリスト）でなく、観察されなかった**種の数**なら、**いくつかの仮定を設ける**ことで、統計的に推定できるのである。そんな馬鹿な、という気分にさせられるが、以下の 3 つのアイデアが、この一見不可能を可能にしてくれる。

1．　観察できなかった種の数はいくつ以上あるという数式を作る。それをもとにデータからいくつ以上という推定法を作る。
2．　観察されなかった種たちで全体の相対頻度の何割を占めるかという数式を作る。それのデータからの推定法を作る。
3．　種数－個体数曲線を伸ばすことで真の種数を推定する。その傾きとデータを結ぶ数式を 1 と 2 を利用して作り、データからの推定法を作る。

革新的で驚嘆させられるのは、1 と 2 である。1 は Chao (1984) で提唱された。2 は、古く第 2 次世界大戦における Turing の研究に遡る。

以下では、この 3 つのアイデアを順に説明し、最後にそれらを統合して、2 章の式(2.1)で定義されるヒル数型の多様性指数 $^{q}D=(\sum_{i=1}^{S}p_i^{q})^{\frac{1}{1-q}}$ の推定式を紹介する。数式がたくさん出てくるが、ほぼすべて大学初年級の 2 項分布や多項分布に関するもので、実質、高校数学の域を出ない。実際のところ、以下

の数式を丹念に追うのは一部の数理統計を専門とする（専門のひとつとしたい）人だけで十分で、一般には、実際の（または人工的に作った）種個体数を用いたシミュレーションで、本当に観察できなかった種をうまく捉えた推定のできることを「実体験」した上で用いていけば十分であろう。

それに先立ち、数学の記号の準備をしておく。

3.1 集団とそのサンプルに関する数学記法の用意

調べている集団には S 種生息し、それらの相対頻度を $p_1, \ldots, p_S$ とする。これらは真値であり、当然、我々にはわからない（そもそも i としてどの S 個の種を挙げればいいのかさえわからない。とりあえず図鑑にある全種か?)。また、推定したいのは、すべての生息種の相対頻度（すべての p_i）ではなく、これを用いて定義されるヒル数型の多様性指数 ${}^qD = (\sum_{i=1}^{S} p_i^q)^{\frac{1}{1-q}}$ の値である。この点を混同してはならない。

この集団から、n 個体のサンプルをとったところ、種 i は x_i 個体、観察されたとする。種 i が観察されなかったら、$x_i = 0$ である。

サンプリングが毎回**独立**で、観察した個体はすぐに集団に戻したとする（**反復を許すサンプリング**)。この 2 つの仮定に違和感を抱かない調査経験者はいないだろう。実際の植物の調査では、1 回確認した個体は重複して数えることのないようラベル付けなどを行う。当然、後者の仮定は満たされない。動物の捕獲調査でも同じわなの中に反復はないし、そもそも同じ個体を何度も確認することのないよう気を付けて調査するのが普通である。前者の独立性の仮定についても、ある個体の近くの個体は観察されやすいし、植物ではある調査区内の全個体を確認するから、やはりたいていの調査において満たされない。

ただ、集団がある程度大きいと、重複を許さない場合でも、許す場合と得られる数値結果はほぼ同じとなる。独立性も、調査区や罠の数を増やすことで、独立とみなせる場合も多くなる。ここではひとまず独立性と重複を許すという仮定での理論的枠組みを示す。これが基本モデルとなる。野外調査の実状に合わせた推定法は、この**基本モデルを順次修正**することで考案していくのが（0 からすべて作り上げるより）賢明である。

独立性と反復を許すサンプリングを仮定すると、種 i の観察数 x_i は、調査のたびに変動する確率変数とみなせる。数学の慣習に従って、確率変数としての数値は大文字、データという現実の数値（やサンプル数 n のような決められた数値や p_i のような真の数値）は小文字で表すと、これは、

$$P(X_1 = x_1, \ldots, X_S = x_S) = \frac{n!}{x_1! \cdots x_S!} \cdot p_1^{x_1} \cdots p_S^{x_S} \quad (3.1)$$

と書ける。

普通に考えると、多様性指数の相対頻度 p_i を観察された相対頻度

$$\hat{p}_i = \frac{x_i}{n} \quad (3.2)$$

で(2.1)を置き換えた式

$$(\sum_{i=1}^{S} \hat{p}_i^{\,q})^{\frac{1}{1-q}} \quad (3.3)$$

で推定してしまい、それで問題ないように思う。実際、多項分布モデルの元での p_i の最尤推定量は、式(3.2)で与えられる。最尤推定量をべき乗したり加えたりするが、最尤推定値を用いた推定値はそれなりに適切なはずである。ただ、繰り返しになるが、それだと観察されなかった種の相対頻度を 0 としてしまう。最も尤もらしい推定値であるが、$p_i > 0$ が真実なのだから、明らかに観察されなかった種の相対頻度を過少推定している。そのため、特に q が 0 や 0 に近い場合、多様性指数(2.1)は、確実に過少推定になっている。

どのような工夫をしたところで、常に真値と等しくなる推定式が作られるわけではない。重要なのは、多くのデータに対しより真値に近い数値になってくれるような推定法の開発である。最尤推定値を用いる推定は、もちろんそれなりに真値に近い。それよりもっと真値に近くなる場合が多い推定式の開発が目標である。

そのような推定法を作るため、いくつか数学の記号を用意しておく。

観察された個体数が k である種の数を f_k で表す。k の範囲は $0 \le k \le n$ である。f_0 は 0 個体観察された、即ち観察されなかった種数を表す。

この記号は多様性の統計数理で頻繁に見かけるが、この業界以外ではあまり見られない記法である。特に、実際のデータでは、特に大きな k で $f_k = 0$ が多い。なぜなら、例えば 100 個体調査し、最も優占した種が 50 個体、2 番目の優占種が 20 個体だったら、$f_{50} = 1$ で、50 から上の 100 までと 21 から 49 までの f_k はすべて 0 となる。こんな文字を増やしてどうするのかと思うのだが、多様性に関する数理では、この記法はしばしば有用である。

f_k は調査によって毎回変化する。だから確率変数として F_k と表すほうが望ましい。確率変数として数式で表すと、indicator 関数

$$I(Z) = \begin{cases} 1 & Z\text{が正しいとき} \\ 0 & Z\text{が正しくないとき} \end{cases} \quad (3.4)$$

を用いて

$$F_k = \sum_{i=1}^{S} I(X_i = k) \qquad (3.5)$$

となる。

相対頻度 p_i の種が k 個体観察される確率は、2 項分布を用いて

$$P(X_i = k) = \binom{n}{k} p_i^k (1-p_i)^{n-k} \qquad (3.6)$$

だから、F_k の期待値は、

$$\mathbf{E}[F_k] = \sum_{i=1}^{S} \mathbf{E}[I(X_i = k)] = \sum_{i=1}^{S} (P(X_i = k)\cdot 1 + P(X_i \neq k)\cdot 0) = \sum_{i=1}^{S} \binom{n}{k} p_i^k (1-p_i)^{n-k} \quad (3.7)$$

となる。ここで、$\binom{n}{k}$ は組み合わせの数

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} = \frac{n!}{k!(n-k)!} \qquad (3.8)$$

である。観察された f_k は $\mathbf{E}[F_k]$ と近いことが予想されるが、等しいわけではない（そもそも期待値は整数とは限らない）。

k 個体観察された種の相対頻度の平均値を a_k とする。数式で書くと

$$a_k = \frac{\sum_{i=1}^{S} p_i I(x_i = k)}{f_k} \qquad (3.9)$$

である。これも確率変数として

$$A_k = \frac{\sum_{i=1}^{S} p_i I(X_i = k)}{\sum_{i=1}^{S} I(X_i = k)} \qquad (3.10)$$

と書くことができる。

3.2 観察できなかった種の数

$\mathbf{E}[F_k]$ に関する式(3.7)で $k = 0, 1, 2$ とすると

$$\mathbf{E}[F_0] = \sum_{i=1}^{S} (1-p_i)^n \qquad (3.11)$$

$$\mathbf{E}[F_1] = \sum_{i=1}^{S} n p_i (1-p_i)^{n-1} \qquad (3.12)$$

$$\mathbf{E}[F_2] = \sum_{i=1}^{S} \frac{n(n-1)}{2} p_i^2 (1-p_i)^{n-2} \qquad (3.13)$$

となる。ここで、シュヴァルツの不等式

$$\sum_i \alpha_i^2 \sum_i \beta_i^2 \geq (\sum_i \alpha_i \beta_i)^2 \quad (3.14)$$

に $\alpha_i = (1-p_i)^{\frac{n}{2}}$、$\beta_i = p_i(1-p_i)^{\frac{n}{2}-1}$ を代入すると、

$$\sum_{i=1}^{S}(1-p_i)^n \sum_{i=1}^{S} p_i^2(1-p_i)^{n-2} \geq (\sum_{i=1}^{S} p_i(1-p_i)^{n-1})^2$$

となる。両辺を $\sum_{i=1}^{S} p_i^2(1-p_i)^{n-2}$ で割ると左辺は上の $\mathbf{E}[F_0]$の式(3.11)に等しくなり、$\sum_{i=1}^{S} p_i(1-p_i)^{n-1} = \frac{\mathbf{E}[F_1]}{n}$、$\sum_{i=1}^{S} p_i^2(1-p_i)^{n-2} = \frac{2\mathbf{E}[F_2]}{n(n-1)}$ から、

$$\mathbf{E}[F_0] = \sum_{i=1}^{S}(1-p_i)^n \geq \frac{(\sum_{i=1}^{S} p_i(1-p_i)^{n-1})^2}{\sum_{i=1}^{S} p_i^2(1-p_i)^{n-2}} = \frac{(\frac{\mathbf{E}[F_1]}{n})^2}{\frac{2\mathbf{E}[F_2]}{n(n-1)}} = \frac{n-1}{n} \cdot \frac{(\mathbf{E}[F_1])^2}{2\mathbf{E}[F_2]} \quad (3.15)$$

という不等式が得られる。

これは、観察されない種数の期待値 $\mathbf{E}[F_0]$が、1 個体及び 2 個体観察される種数の期待値 $\mathbf{E}[F_1]$と $\mathbf{E}[F_2]$で下から抑えられるということを意味する。

そこで、この中の期待値を観察値で置き換えた

$$\hat{f}_0 = \frac{n-1}{n} \cdot \frac{f_1^2}{2f_2} \quad (3.16)$$

でもって観察されなかった種数 F_0 の推定値としたのが、Chao (1984)で提唱されたものである。上の通り、基本的には観察されなかった種数はこれ以上あるだろうという推定値である。また、期待値と観察値は近いとはいえ実際には異なるので、これで「正しく」下からの評価を与えているわけではない。あくまで、これ以上いるだろうという目安のひとつである。

観察されなかった種については無情報なのだが、それが何種くらい以上あるかなら、1 個体だけ観察された種 (singleton という) 及び 2 個体観察された種に関する情報から推定できるという、画期的な数式である。

この推定法が提唱されて以来、様々な実データや人工データの種数と相対頻度を元にこの推定法の妥当性が調べられてきている。今日まで、観察されなかった種数の推定として頻繁に利用され、その有益性が実証されている。

3.3 観察されなかった種全体が占める割合

Chao (1984)による観察されなかった種数 F_0 に関する下からの評価に加え、観察されなかった種の相対頻度 a_0 の合計（全体における割合）も推定可能で

ある。

F_kの定義式(3.5)に出てくる和$\sum_{i=1}^{S} I(X_i = k)$でなく、各項に$\frac{p_i}{1-p_i}$をかけた確率変数の期待値を考える。

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

(3.7)と同じように期待値を確率に直すと

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

分子分母に $k+1$ と $n-k$ をかけると、

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

シグマの部分はちょうど式(3.7)で k の代わりに k + 1 とした式と同じだから$\mathbf{E}[F_{k+1}]$である。従って、

$$\mathbf{E}[\sum_{i=1}^{S} \frac{p_i}{1-p_i} I(X_i = k)] = \frac{k+1}{n-k} \mathbf{E}[F_{k+1}] \quad (3.17)$$

を得る。

ここで、考えている確率変数 (X_i) 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$\frac{r+1}{n-r} \cdot f_{k+1}$に近い値になる。

一方、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。そして$\sum_{i=1}^{S} I(X_i = k)$は f_k となる。従って、左辺は$\frac{a_k}{1-a_k} \cdot f_k$に近いはずである。

まとめると、

$$\frac{r+1}{n-r} \cdot f_{k+1} \approx \frac{a_k}{1-a_k} \cdot f_k \quad (3.18)$$

という式が得られた。ここで ≈ は数値が近い（近似的に成り立つ）ことを意味する。

これをa_kについて解くと、

$$\hat{a}_k = \frac{(r+1) f_{k+1}}{(n-r) f_k + (r+1) f_{k+1}} \quad (3.19)$$

という推定式が導かれる。

(3.19)で $k = 0$ とすると、

$$\hat{a}_0 = \frac{f_1}{nf_0 + f_1} \quad (3.20)$$

となる。$k \geq 1$ なら観察されたf_kがあるが、f_0についてはこれがない。だから、(3.20)の右辺は計算できない。これは、本来、観察されなかった種については無情報なので、a_0は推定できなくて当然の帰結である。

そこで、f_0の代わりに、3.2 節で出てきた Chao (1984)で提唱された、（本来は下からの）推定値$\hat{f}_0 = \frac{n-1}{n} \cdot \frac{f_1^2}{2f_2}$を用いると、

$$\hat{a}_0 = \frac{2f_2}{(n-1)f_1 + 2f_2} \quad (3.21)$$

という推定式を得る。ここでも、式(3.16)のときと同様、観察されなかった種に関する推定が、1 個体及び 2 個体観察された種に関する情報で可能となっている。

実際に有用なのは、a_0 という観察できなかった種の平均相対頻度より、それらが全体で何割を占めるかという、$\sum_{i=1}^{S} p_i I(x_i = 0) = a_0 f_0$ のほうである。これは直観的には、手にしたデータで集団全体の何パーセントを把握していないかの、ひとつの定量化である。逆に見ると、$\sum_{i=1}^{S} p_i I(x_i > 0) = 1 - a_0 f_0$は、手にしたデータで集団の何パーセントを抑えられたかと解釈できる。そこでこれを、確率変数の言葉で

$$C_n = \sum_{i=1}^{S} p_i I(X_i > 0) = 1 - \sum_{i=1}^{S} p_i I(X_i = 0) \quad (3.22)$$

で定義し、サンプル数 n のデータにおけるサンプル被覆度 (sample coverage)という。定義から

$C_n = 1 - A_0 F_0$ (3.23)

が成り立つ。また、この期待値は

$$\mathbf{E}[C_n] = 1 - \mathbf{E}[\sum_{i=1}^{S} p_i I(X_i = 0)] = 1 - \sum_{i=1}^{S} p_i (1 - p_i)^n \quad (3.24)$$

となっている。

a_0 f_0の推定値として、(3.21)に f_0をかけ、さらにf_0の代わりに$\hat{f}_0$を用いると、

$$\hat{a}_0 \hat{f}_0 = \frac{2f_2}{(n-1)f_1 + 2f_2} \frac{n-1}{n} \frac{f_1^2}{2f_2} = \frac{f_1}{n} \cdot \frac{(n-1)f_1}{(n-1)f_1 + 2f_2} \quad (3.25)$$

が得られた。(3.23)より、1 からこれを引いたものは、サンプル被覆度 C_n の推定値として使える。

$$\hat{C}_n = 1 - \frac{f_1}{n} \cdot \frac{(n-1)f_1}{(n-1)f_1 + 2f_2} \quad (3.26)$$

概して、労して種個体数データを集めても、それで集団全体の何パーセントを抑えられたのか、常に不安がつきまとうものである。集団の真の構成を知らないのだから、自分が今どの位置にいるか知りようもないと思うのが自然だが、手元のデータから推定可能なのである。

$\hat{a}_0\hat{f}_0$ の推定式(3.25)も、不思議な式である。右辺は、n が大きいと後半はほぼ 1 となるので、ほぼ $\frac{f_1}{n}$ である。

$$\hat{a}_0\hat{f}_0 \approx \frac{f_1}{n} \quad (3.27)$$

これは、1 個体だけ観察された種の数 f_1 をサンプル数 n で割っている。1 個体だけ観察された種の相対頻度を普通に 1/ n で推定すると、それらが f_1 種あるので全体で占める割合は $\frac{f_1}{n}$ となる。ところが、式(3.27)は、これは観察されなかった個体の占める割合の推定だと主張するのである。

では、1 個体だけの種が占める割合はどう推定するかというと、式(3.19)で $k = 1$ とし、両辺に f_1 をかけて

$$\hat{a}_1 f_1 = \frac{2f_1 f_2}{(n-1)f_1 + 2f_2} \quad (3.28)$$

という式を用いる。今度は、2 個体だけ観察された種の数 f_2 を含む式となっている。

信じがたい結果であるし、(3.25)を導く段階で $\hat{a}_0$ と $\hat{f}_0$ の近似を使っているので、必ずしもこうした推定のほうが最尤推定値 $\hat{p}_i = x_i / n$ を用いる推定より優っているという保証はない。しかし、数多くのシミュレーション結果が、(3.25)や(3.28)の優越性を示している。最尤推定値が万能とは限らない好例といえよう。

3.4 種数－個体数曲線の傾き

こうして、観察されなかった種数や、手持ちのデータが把握しているのが集団全体の何パーセントくらいか（サンプル被覆度）という、常識的には推

定できるはずのない数量が、手持ちのデータから推定できることがわかった。次は、ヒル数型の多様性指数(2.1)の推定式の作成である。単純に相対頻度をデータから最尤推定した $\hat{p}_i = x_i / n$ を用いるより優れた推定式を作るには、どうすればいいだろう。

ここでは、種数－個体数曲線の傾きを用いる推定法を紹介する。

種数－個体数曲線とは、横軸に個体数 k、縦軸にそのサンプル数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ものである。サンプル数 n の実際のデータがあるときは、その中の k 個を用いることで $S(k)$を推定し、この曲線を描くことができる。なお、常に $S(0) = 0, S(1) = 1$ が成り立つ。

こうして描かれる種数－個体数曲線は、多くの場合、単調に増加するが増加はしだいに頭打ちとなり、どこかに収束するような様相を示す。だから曲線を適当に延長して収束したときの縦軸を読めば、それが真の種数のはずである。ただ、どう曲線を延長するかが難しい。例えば、何らかの数式を作り、最小 2 乗法などでパラメータを最適化して求めるという手段が考えられる。そのとき、考えた数式に特に根拠がなく当てはまりが良さそうだけというなら、とりわけ数式の形によって推定種数が異なっている場合、推定値に信憑性が伴わない。また、どのくらい数式が曲線と近かったらいい推定と言えるのかについて根拠ある目安がないと、やはり信憑性に欠ける。

曲線全体でなく、傾きに注目するとどうなるだろう。傾きといっても、横軸の個体数は整数値しかとらないから、隣り合う数値の差をとるだけである。これを $\Delta(k)$ と書くことにする。

$$\Delta(k) = \frac{S(k+1) - S(k)}{(k+1) - k} = S(k+1) - S(k) \quad (3.29)$$

さて、冒頭の多項分布モデルのもとでは、$S(k)$はサンプルに依存して変動する確率変数となる。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は $(1-p_i)^k$ だから、1 度は観察される確率は $1-(1-p_i)^k$ である。どの種も、観察されれば 1 と数えられ、観察されなければ 0 と数えられるから、$S(k)$の期待値は

$$\mathbf{E}[S(k)] = \sum_{i=1}^{S} \{(1-(1-p_i)^k)\cdot 1 + (1-p_i)^k \cdot 0\} = \sum_{i=1}^{S} 1 + \sum_{i=1}^{S} (1-p_i)^k = S - \sum_{i=1}^{S} (1-p_i)^k \quad (3.30)$$

である。

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

だから、

$$\mathbf{E}[\Delta(k)] = \mathbf{E}[S(k+1)] - \mathbf{E}[S(k)] = \sum_{i=1}^{S} p_i (1-p_i)^k \quad (3.31)$$

となる。なお、常に $S(0)=0, S(1)=1$ だから、常に $\Delta(0)=1$ が成り立つ。なお、(3.31)を 1 から引いたものは、サンプル被覆度 C_k の期待値(3.24)と一致している。

Hill 型の多様性指数で必要な計算は、$\sum_{i=1}^{S} p_i^{\,q}$ である。これを

$$^qD = \sum_{i=1}^{S} p_i^{\,q} \quad (3.32)$$

とおき、^qD と $\mathbf{E}[\Delta(k)]$(3.31)の関係式を作る。まず、

$$^qD = \sum_{i=1}^{S} p_i^{\,q} = \sum_{i=1}^{S} p_i (1-(1-p_i))^{q-1}$$

と書き、テーラー展開（2 項分布の一般の実数べきへの拡張）

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

で $x=1$ としたものを用いると、

$$^qD = \sum_{i=1}^{S} p_i \sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k (1-p_i)^k = \sum_{i}^{} \sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k)$$

この無限和を、サンプル数 n の一つ手前と n 以降分ける。

$$^qD = \sum_{k=0}^{n-1} \binom{q-1}{k} (-1)^k \Delta(k) + \sum_{k=n}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k) \quad (3.33)$$

第 1 項の有限和の中の$\Delta(k)$について、まず、次の命題を示す。

命題 3.1

$k < n$ のとき、確率変数 $\dfrac{X_i}{n} \dfrac{\binom{n-X_i}{k}}{\binom{n-1}{k}}$ の期待値は $p_i(1-p_i)^k$ である。

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

証明

$P(X_i = x) = \binom{n}{x} p_i^{\,x} (1-p_i)^{n-x}$ だから、

$$\mathbf{E}[\frac{X_i}{n}\frac{\binom{n-X_i}{k}}{\binom{n-1}{k}}]=\sum_{x=0}^{n}\binom{n}{x}{p_i}^x(1-p_i)^{n-x}\frac{x}{n}\frac{\binom{n-x}{k}}{\binom{n-1}{k}}$$

$x>n-k$ だと $k>n-x$ で、$\binom{n-x}{k}=\frac{(n-x)(n-x-1)\cdots(n-x-(k-1))}{k(k-1)\cdots1}$ の分子は $n-x$ から 1, 2, …, $k-2$, $k-1$ を順に引いているがその中の一つはちょうど $n-x$ になるので 0 になる。つまり、シグマは実質的には $x=0$ から $x=n-k$ で済む。そこで $p_i(1-p_i)^k$ を前に出し、組み合わせで表される階乗を書き下すと、

$$=p_i(1-p_i)^k\sum_{x=1}^{n-k}{p_i}^{x-1}(1-p_i)^{n-k-x}\frac{n(n-1)\cdots(n-x+1)}{x!}\frac{x}{n}\frac{(n-x)\cdots(n-x-k+1)}{(n-1)(n-2)\cdots(n-1-k+1)}$$

分子を整理してよく見ると、以下のように n から $n-x-k+1$ までつながっていることに気づく。

$$=p_i(1-p_i)^k\sum_{x=1}^{n-k}{p_i}^{x-1}(1-p_i)^{n-k-x}\frac{n(n-1)\cdots(n-x+1)}{(x-1)!}\frac{(n-x)\cdots(n-k)(n-k-1)\cdots(n-x-k+1)}{n(n-1)(n-2)\cdots(n-k)}$$

分母とキャンセルした残りは

$$=p_i(1-p_i)^k\sum_{x=1}^{n}\frac{(n-k-1)\cdots(n-k-1-(x-1)+1)}{(x-1)!}{p_i}^{x-1}(1-p_i)^{n-k-x}$$

となるが、分数の部分はちょうど $n-k+1$ から $x-1$ 取り出すときの組み合わせの数と一致している。

$$=p_i(1-p_i)^k\sum_{x=1}^{n-k}\binom{n-k-1}{x-1}{p_i}^{x-1}(1-p_i)^{n-k-x}$$

シグマの部分は、$p_i+(1-p_i)$ の 2 項展開の形となっているから 1 になるので、

$$=p_i(1-p_i)^k$$

となって証明できた。

この命題の右辺を i についてとった和が$\Delta(k)$の期待値と等しいので、左辺を i について和をとったものは $\mathbf{E}[\Delta(k)]$の不偏推定値となっている。そこでこれを $k<n$ のときの$\Delta(k)$の推定値に使う。

$$\hat{\Lambda}(k)=\sum_{i=1}^{S}\frac{x_i}{n}\frac{\binom{n-x_i}{k}}{\binom{n-1}{k}}\quad(k<n)\qquad(3.34)$$

こうして、${}^qD$ の推定式(3.33)の第 1 項の推定量を得られた。

後半では、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。それは、$S(n+m)$ 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相対頻度は(3.9)で $k=0$ とした 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 $1-(1-a_0)^m$ である。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、$f_0(1-(1-a_0)^m)$ である。従って、サンプル数 n のデータ $(x_1,\ldots,x_S)$ の元での $S(n+m)$の条件付き期待値は、サンプル数 n のデータの中で観察された種数を S_{obs} と書くと、

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

となる。

同じように、サンプル数 n のデータの元での $S(n+m+1)$ の条件付き期待値は、

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

だから、サンプル数 n のデータの元での $\Delta(n+m) = S(n+m+1) - S(n+m)$ という確率変数の条件付き期待値の推定値として、下から上を引いた

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

を得る。

従って、$\Delta(n+m)$ の推定値として、ここに a_0 と f_0 の推定値 $\hat{a}_0$(3.21)と $\hat{f}_0$(3.16)を代入した、

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

が得られる。なお、 (3.21)と(3.16)から $\hat{a}_0\hat{f}_0 = \frac{f_1}{n}\cdot\frac{(n-1)f_1}{(n-1)f_1+2f_2} = \frac{f_1}{n}(1-\hat{a}_0)$ となるから、

$$\hat{\Delta}(n+m) = \frac{f_1}{n}(1-\hat{a}_0)^{m+1} = \frac{f_1}{n}\left(1-\frac{2f_2}{(n-1)f_1+2f_2}\right)^{m+1} \qquad (3.35)$$

と書ける。

(3.34)と(3.35)を(3.33)に代入したものが ${}^qD$ の推定量である。

$${}^q\hat{D} = \sum_{k=0}^{n-1}\binom{q-1}{k}(-1)^k\hat{\Delta}(k) + \sum_{m=0}^{\infty}\binom{q-1}{n+m}(-1)^{n+m}\hat{\Delta}(n+m) \quad (3.36)$$

q が 0, 1, 2 の場合、(3.36)は簡単な形にまとめられる。

$q = 0$ のとき、$\binom{-1}{k} = \frac{(-1)(-2)\cdots(-k)}{k!} = (-1)^k$ だから、

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

第 1 項について、以下の命題が成り立つ。

命題 3.2

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

証明

まず、組み合わせを書き下すことで

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

がわかる。次に、

$$\sum_{k=0}^{n-1}\hat{\Lambda}(k) = \sum_{k=0}^{n-1}\sum_{1\le X_i\le n-k}\frac{\binom{n-k-1}{x_i-1}}{\binom{n}{x_i}} = \sum_{1\le x_i\le n}\sum_{k=0}^{n-x_i}\frac{\binom{n-k-1}{x_i-1}}{\binom{n}{x_i}}$$

とシグマの順序を交換し、n に関する数学的帰納法を使えば、

$$\sum_{k=0}^{n-x_i}\binom{n-k-1}{x_i-1} = \binom{n}{x_i}$$

が示されるため、

$$\sum_{k=0}^{n-1}\hat{\Lambda}(k) = \sum_{1\le x_i\le n}1$$

となり、x_i は観察された種数 S_{obs} だけ動くから、命題は示された。

第 2 項の無限和は、等比級数の和の公式を用いて

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

これは、f_0 の推定値 $\hat{f}_0$ (3.16)にほかならない。結局、

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

となった。何だかんだで、観察された種数に、観察されていない種数の（下からの）推定量を加えた式となった。

参考文献

Chao A. (1984) Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11: 265-270.

Chao A and Jost L. (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93: 2533-2547.

Chiu C.H., Wang Y.T., Walther B.A. and Chao A. (2014) An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. Biometrics 70: 671-682.

Chao A. and Jost L. (2015) Estimating diversity and entropy profiles via discovery rates of new species. Methonds in Ecology ad Evolution. Doi: 10.1111/2041-210X.12349.