中立遺伝マーカーを用いた近親判別に基づく個体数推定の可能性

入　江　貴　博

Stock size estimation based on the close-kin relationship detected from neutral genetic markers

Takahiro Irie

いりえ たかひろ：東京大学大気海洋研究所

中立遺伝子情報に基づいて判明した親子ペア数から，野外での産卵親魚個体数を推定する，資源量推定のためのクロスキン分析の方法を概説する．単純化された仮定の下では，親子ペア数は超幾何分布に従うことを示した上で，逆に観測された親子ペア数から個体数をベイズ推定する手順の数学的な背景を簡単に紹介する．

1. クロスキン法：その開発の背景

国内外を問わず，水産資源評価のための従来の資源量推定は，漁業者によってなされた漁獲に関するデータを，VPA や統合モデルなどの資源評価モデルに入力することで実施されてきた．タグ調査，調査船による加入量調査，航空機による目視調査といった資源量推定のためだけに収集されたデータを加味することも多いが，計算に用いられる情報の根幹は漁業に由来するデータである．当然のことであるが，漁業者は，個体群から魚を無作為抽出しているわけではない．その一方で，統計学的推定の学術体系に依拠した資源評価モデルは，無作為抽出されたデータを計算の対象とすることで初めて，最も正しい資源量を推定することができる．資源評価モデルには，漁獲データに内在する非無作為抽出性を補正する枠組みも組み込まれてはいるものの，その効果には限度がある．そのような背景から，漁業から独立した資源量推定のための指標を確立することの必要性が，資源評価の現場では叫ばれ続けてきたのであった．

DNA シーケンシングに関する近年の技術的発展は，たいへん目覚ましい．来たるべき技術的飛躍を見越して，遺伝情報を水産資源の個体数推定に用いるための模索は，古くから行われてきた．第二章で荒木仁志博士が紹介されている，中立遺伝マーカーを用いた有効集団サイズの推定もその一例である．本稿で紹介するクロスキン法は，筆者が知る限りでは，豪州の研究者 Mark Bravington 博士がその開発と普及に献身を続けてきた技術である．資源量推定のための実用化に向けて，まずは大型鯨類 [1] とミナミマグロ [2] への適用が試みられた．現在は，これらと並行して，タイヘイヨウ

クロマグロでの実施の準備が進められている．

2. クロスキン法の概略

実在する生物の個体群の構造は，空間的にも時間的にもたいへん複雑である．生活史は複数の個体発生段階からなり，被食や斃死や漁獲による累積死亡で，個体数は加齢とともに減少する．海流によって受動移送されたり，自らの遊泳能力によって能動的に移動することで，空間的にも均質ではない分布パターンを作り出す．また，個体群サイズは大規模な年々変動を繰り返すことが常である．そのように複雑な系を忠実に再現するようなモデルをはじめから構築しようとすると，人間の脳には荷が重すぎて，計画は途中で頓挫してしまうことだろう．従って，ここでは問題解決の第一歩として，数学的に扱いやすい，最も単純な状況を想定した上で，中立遺伝マーカーの情報から個体群サイズを推定するための数理モデルを導出してみる．

ここで想定するのは，(1) 雌が次の世代に必ず 1 個体の子を残す（雄はいない），(2) 閉鎖個体群で外部との移出入はない，(3) 世代の重複はない，という現実には到底ありえない生活史を示す生物である．親世代の個体は，ある時点で一斉に子を 1 個体残し，その次の瞬間には死亡する．残された子世代の個体は，孫世代の個体を残すまでは 1 個体たりとも死亡しない．以上の設定は著しく現実離れしているのだが，数理モデルの簡略化という点ではこれが最も好都合な仮定である．

生活史と個体群動態が決まったので，次に個体数と標本数を決める．ここでは親世代と子世代という二世代を取り扱う．親世代の総個体数を N，子世代の総個体数を M としよう（図 1）．この生物は子を残すまで死なないので，個体群サイズは 0 でない限りはいつでも N もしくは M である．そして，1 親が必ず 1 子を残すので，$N = M$ である．次に標本数であるが，親子それぞれの個体群から，無作為に n と m の数の標本を採集することにする（図 1）．親の採集は，子を残した直後に行うので，採集によって子を残す親の数が減るという心配はご無用である．

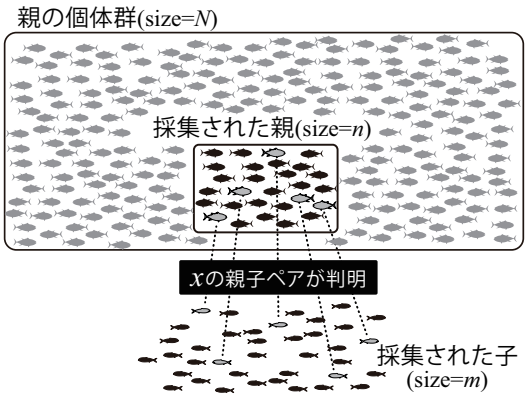

図 1　総数 N の親個体群と総数 M の子個体群からそれぞれ n 個体と m 個体を無作為抽出した結果，x の親子ペアが判明した．

無作為抽出した n 個体と m 個体の標本の遺伝情報を調べて，親子関係にあるペアの数を決めよう（具体的にどう調べて，どう親子の判別をするかについては，次節で簡単に紹介する）．見出された親子ペアのことを POP と呼ぶ（Parent-Offspring Pair の略）．ここで，DNA を調べた結果，x 組の親子ペアが見出されたとする（図 1）．さて，親世代から無作為抽出された n 個体と子世代から無作為抽出された m 個体の間に x 組の親子ペアが見出された場合，親世代の個体数 N はいくつだと推定されるか？クロスキン分析を用いた資源量推定というのは，つまるところ，こういう問いなのである．

深遠なる統計学的推定の理論やモデルを持ち出す前に，とりあえず直感だけで考えてみよう．親の標本数 n が 100，子の標本数 m が 100，見つかった親子ペア数 x が 5 だったとする．採集された子のうち 5 個体には，標本中に親が見つかったわけだが，残りの子 95 個体の親は，採集されなかった親の個体群にいることになる．100 個体の親を調べると，そのうち 5% の個体に子が見つかるわけだから，すべての子の親が見つかるように親の標本数を増やすためには，親の標本数を 20 倍しなければならない．この 20 倍した親の数というのは，すなわち親の全個体数である（調べたすべての

子の親を見つけるためには，すべての親を調べる必要がある）．以上の説明を数式で示すと，次のようになる：

$$N = n \times (m/x) = 100 \times (100/5) = 2000 \qquad (1)$$

この $N = nm/x$ という式は，標識再捕の分野で用いられている Lincoln-Petersen 推定量と同じものである [3]．ここで仮定している一子相伝の生命体に限っては，親と子の立場を入れ替えても計算が成立する．つまり，N と M，n と m をそれぞれ入れ替えても式が成り立つということである．実際の生物では，子の数はまちまちであったり，死亡があったりして，子の数の推定は親の数の推定よりも難しそうである．

3. DNA を用いた親子判別の方法

ここまで，意味を説明せずに「クロスキン分析」という語を用いてきた．クロスキンは，close-kinship という英単語からできた造語で，近親関係とか近い血縁関係というような意味である．野外で採集した同種個体が，互いに親子関係あるいは兄弟関係にあるかどうかを判定するためには，DNA に刻まれた遺伝情報を解析する必要がある．本節ではその原理を説明するが，理解のための準備として，高校の生物の授業で習ったメンデルの法則を思い出していただきたい．

前節では，雌しかいない仮想生物にご登場いただいたが，ここでは雌雄異体で有性生殖を行う二倍体の生物を考える．この場合，よく引き合いに出されるのは 1 遺伝子座 2 対立遺伝子モデルである．ある遺伝子座が a と b というアレル（対立遺伝子）を持つ場合，この遺伝子座に関する父親の遺伝子型は aa，ab，bb のいずれかである．母親も同様．この両親から生まれた子の遺伝子型は，両親の遺伝子型によって決まる．ここで，両親の遺伝子型によっては，子が持ち得ない遺伝子型が存在するという点が重要である．たとえば両親が ab ならば，aa，ab，bb すべての子が生まれ得る（図 2A）．父親が aa で母親が bb ならば，子はかならず ab となる（図 2B）．両親がともに aa という遺伝子型の場合，子の遺伝子型は必ず aa となる

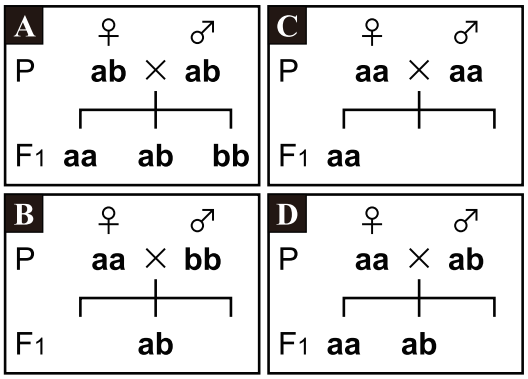

図 2　親 (P) の遺伝子型から生じうる子 (F_1) の遺伝子型．

（図 2C）．父親が aa で母親が ab ならば，bb という子は生まれない（図 2D）．つまり，子の遺伝子型を見ることで，親ではあり得ない個体を親候補からはじくことができる．実際には，1 遺伝子座からの情報だけでは，親の候補が絞り込めないため，親子の判別はできない．そこで，遺伝子座の数をたくさん増やすことで，親でない個体を親であると判断してしまう過誤の確率を 0 に近づけていく．このような親子判別の方法を，排除法という．

この方法で親子関係を判別した場合，特定された親の確からしさは，子の遺伝子型によって異なるという事実がある．ヘテロ接合（ab）となる遺伝子座をたくさん持つ子ほど，絞り込まれた親が真の親であるという確からしさは，低くなる．それから排除法では，考慮する遺伝子座の数を増やせば増やすほど，真の親を親でないと判断してしまう過誤の確率が上がってしまう．実際の遺伝情報には，突然変異や塩基配列決定時のミスなどから生じた情報の改変が，わずかではあるが含まれるためである．このような親子判別の確からしさのムラを量的に考慮したい場合は，上の排除法ではなく，尤度を用いた方法（尤度法など）を導入する必要がある．紙面の制約から，本稿では尤度を用いた方法を紹介することはできないが，実際の資源量推定では，排除法よりも尤度に基づく方法を用いることが好ましいだろう．クロスキン

式(3)

$$
\begin{aligned}
\mathrm{P}(x|N=M,n,m) &= \frac{\begin{bmatrix}\text{ペアを作る}\\ \text{親の選び方}\end{bmatrix}\times\begin{bmatrix}\text{ペアを作らない}\\ \text{親の選び方}\end{bmatrix}\times\begin{bmatrix}\text{ペアを作らない}\\ \text{子の選び方}\end{bmatrix}}{\begin{bmatrix}\text{親の選び方の}\\ \text{組み合わせ総数}\end{bmatrix}\times\begin{bmatrix}\text{子の選び方の}\\ \text{組み合わせ総数}\end{bmatrix}} \\
&= \frac{{}_N\mathrm{C}_x \times {}_{N-x}\mathrm{C}_{n-x} \times {}_{M-n}\mathrm{C}_{m-x}}{{}_N\mathrm{C}_n \times {}_M\mathrm{C}_m} = \frac{{}_N\mathrm{C}_x \times {}_{N-x}\mathrm{C}_{n-x} \times {}_{N-n}\mathrm{C}_{m-x}}{{}_N\mathrm{C}_n \times {}_N\mathrm{C}_m} \\
&= \frac{n!\,(N-n)!\,m!\,(N-m)!}{x!\,N!\,(n-x)!\,(m-x)!\,(N-n-m+x)!} = \frac{{}_n\mathrm{C}_x \times {}_{N-n}\mathrm{C}_{m-x}}{{}_N\mathrm{C}_m} \\
&= \mathrm{HG}(x|N,n,m)
\end{aligned}
$$

分析に基づいて推定された資源量推定値の精度を量的に表現するためには（つまり区間推定を誠実に行うためには），親子判別に伴う不確実性も定量化する必要があるためである．

4. POP が従う確率分布

第 2 節で仮定した一子相伝の生命体に，ここでふたたび登場してもらおう．N 個体の親個体群から無作為に n 個体を抽出し，M 個体の親個体群から無作為に m 個体を抽出する．このとき，親子のペア数すなわち POP の数 x はいくつになるだろうか．式 (1) で与えた $N = n \times (m/x)$ を変形して，次式のようになると答えるかもしれない：

$$x = nm/N \tag{2}$$

しかしながら，実際には POP 数がいつも nm/N になることはないだろう．無作為抽出に際して，確率的なゆらぎ（サンプリングエラー）が生じるためである．つまり，POP 数 x は，確率的にふるまうはずだ．では，POP 数 x の確率的ふるまいは，どのような確率質量関数で記述されるであろうか．

二組のトランプを用意する．ジョーカーは除くので，それぞれ 52 枚のカードからなる ($N = M = 52$)．ひと組から n 枚，もうひと組から m 枚を無作為に抽出したとき，一致する札の数を x としよう．さて，x の確率的ふるまいは，どのような確率質量関数で記述されるだろうか．結論からいうと，札の一致枚数 x は，超幾何分布という離散分布に従う．パラメータは N，M，n，m の 4 個だが，この場合は $N = M$ なので，実質的には 3 個である．そして，ここでいう札の一致枚数は，一子相伝の生命体の例における POP 数で置き換えて考えることができる．つまり，POP 数 x は超幾何分布に従う（ただし，実際の生物で POP 数が超幾何分布に従うことは考えづらい．あくまで一子相伝を仮定した場合の話である）．

細かい説明は省略するが，ここでいう POP 数が超幾何分布 HG (x | N, n, m) に従うという結論は，次式のように導出することができる：式 (3)
実は，超幾何分布 HG (x | N, n, m) に従う確率変数 x の期待値は，$\bar{x} = nm/N$ であることが知られている．先ほど式 (2) で提示した量と右辺同士が一致している点に，ご留意いただきたい．

5. 超幾何分布に基づく親魚個体数の推定

前節では，ある条件の下で，親個体数 N，親の標本数 n，子の標本数 m が決まると，親子のペア数 x の確率的ふるまいが超幾何分布 HG (x | N, n, m) によって記述されることを示した．しかしながら，我々が知りたいのは POP 数 x ではなく，親個体数 N のほうである．POP 数 x は親子の遺伝情報を比べて，定数としての情報を得ているはずである．つまり，POP 数 x，親の標本数 n，子の標本数 m が与えられたときに，親個体数 N がどうなるかを知りたいのである．

こういった条件付き確率の取り扱いに便利な統計学の体系に，ベイズ統計学がある．上で導いた

$$\pi(N|x,n,m) = \frac{{}_n\mathrm{C}_x \times {}_{N-n}\mathrm{C}_{m-x} \times ({}_N\mathrm{C}_m)^{-1} \times (N_{\max}-n+1)^{-1}}{\sum_{N=n}^{N_{\max}} {}_n\mathrm{C}_x \times {}_{N-n}\mathrm{C}_{m-x} \times ({}_N\mathrm{C}_m)^{-1}(N_{\max}-n+1)^{-1}} = \frac{{}_{N-n}\mathrm{C}_{m-x} / {}_N\mathrm{C}_m}{\sum_{N=n}^{N_{\max}} {}_{N-n}\mathrm{C}_{m-x} / {}_N\mathrm{C}_m}$$

式(6)

超幾何分布 HG (x | N, n, m) では，x が確率変数，N, n, m が母数（パラメータ）であった．POP 数 x は，サンプリングのたびに異なる値となり，その挙動は明らかに確率的である．それに対して，3 個の母数は本来ひとつの決まった値を取るものであって，確率的な性質はないように思われる．これに対して，クロスキン分析では，超幾何分布 HG (x | N, n, m) の x と N を入れ替えるようなことをして，観測から得られた POP 数がひとつの値として決まったときに，親個体数 N がどのような値を取りそうかを確率的な視点で推定する．本来確率的でない量を確率的に取り扱うことになるので，これはベイズ統計学がその真価を発揮する問題であると言えるだろう．

さて，本題の親個体数 N の確率分布に話を戻そう．ベイズの定理を用いると，親の標本数 n，子の標本数 m で，判明した POP 数が x であるという条件の下での親個体数 N のふるまいを，事後確率として計算することができるようになる．具体的には，N に関する事後確率 $\pi(N|x, n, m)$ は，x に関する尤度関数 $L(x|N, n, m)$ と N に関する事前確率 $\pi(N)$ を用いて，次式で与えられる：

$$\pi(N|x,n,m) = \frac{L(x|N,n,m)\pi(N)}{\sum_N L(x|N,n,m)\pi(N)} \quad (4)$$

親個体数 N の尤度関数は，実は式 (3) をそのまま使うことができる：

$$L(x|N,n,m) = \frac{{}_n\mathrm{C}_x \times {}_{N-n}\mathrm{C}_{m-x}}{{}_N\mathrm{C}_m} \quad (5)$$

厄介なのは事前分布の選択なのだが，これについては参照できる研究が少なからずある．実は標識再捕法を用いて個体数の推定をする場合に，標識のある再捕個体数もまた超幾何分布に従うため，その際にどんな事前分布を置くべきかという研究が既になされているのだ [4,5]．典型的な事前分布のひとつは，N が n から $N_{\max}$ までの整数を取る確率が等しく，それ以外の値を取る確率は 0 であるという離散一様分布である．もうひとつの典型的な事前分布は，N が 0 から∞までの整数を取る確率がすべて等しいことを仮定した非正則事前分布であるが，こちらは N について 0 から∞まで積分した値が 1 にならない（発散する）．前者の事前分布を仮定した場合，式 (4) は次のように書き直すことができる：式 (6)

事後分布の一例を図 3 に示した．

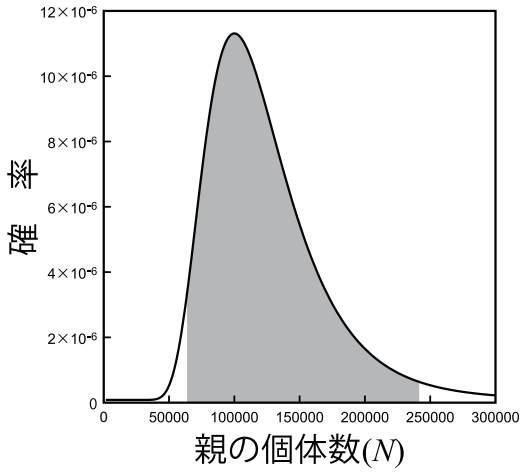

図 3　親の個体数(N)の事後分布. 灰色の領域は 95%信用区間を示す. x=10, n=1,000, m=1,000.

親の総数 N = 100,000，親の標本数 n = 1,000，子の標本数 m = 1,000 を仮定した際に，超幾何分布から計算される POP 数 x の生起確率と，各 POP 数 x の下で式 (6) から計算される親の総数 N の 95% 信用区間 (credible interval) を一覧にしたものが表 1 である．この 95% 信用区間は，親の総数 N が 95%

POP 数(x)	生起確率	N の MAP 推定値	95%CI 下限	95%CI 上限
3	0.007	333333	179765	3759650
4	0.018	250000	138856	1607160
5	0.037	200000	114490	913917
6	0.062	166666	98003	591909
7	0.090	142857	86095	449596
8	0.113	125000	76959	353244
9	0.126	111111	69711	287907
10	0.126	100000	63804	241561
11	0.115	90909	58887	207241
12	0.096	83333	54723	180926
13	0.073	76923	51147	160180
14	0.052	71428	48038	143446
15	0.034	66666	45308	129693
16	0.021	62499	42891	118208
17	0.012	58823	40733	108487
18	0.007	55555	38794	100161
19	0.004	52631	37041	92958
20	0.002	50000	35449	86671
21	0.001	47619	33995	81138
22	0.000	45454	32662	76235
23	0.000	43478	31434	71692
24	0.000	41666	30301	67876

表 1　超幾何分布から計算されるPOP数 xの生起確率 (N=100,000; n=1,000; m=1,000)と，各POP数 xの下で計算された親の総数 Nの事後分布から求めた 95%信用区間 (n=1,000; m=1,000).

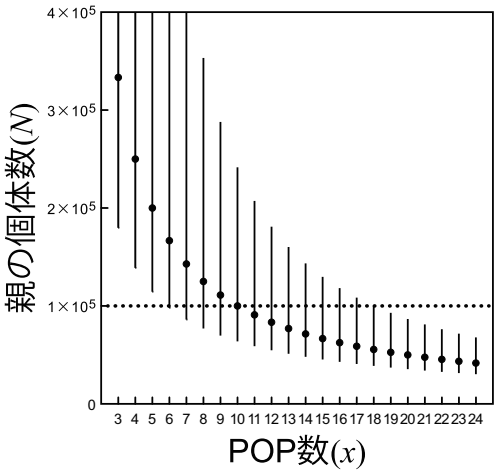

図 4　親の個体数(N)の事後分布から計算した 95%信用区間 (実線) とMAP推定値 (黒点) をPOP数(x)に対してプロットした. 点線は親の個体数(N)の真値(N=100,000). n=1,000, m=1,000.

の確率で含まれる区間推定値を意味する（頻度主義統計学における信頼区間confidence intervalとの意味の違いに注意）. 表 1 の内容をグラフにしたものが図 4 である.

数値計算の結果を示して終わるのでは能が無いので，最後に解析的な成果も少しだけ紹介する. 一様分布のように平坦な事前分布を仮定した場合，N の最大事後確率推定値（MAP 推定値 ; Maximum A Posteriori estimates）は，尤度関数の N に関する偏微分を 0 と等号で結ぶことで，次のように計算できる（途中でスターリングの近似を利用）: 式 (7)

N の MAP 推定値は式 (7) の最後の等式を満たすが，閉形式とはならなかった．N が整数になるようにこの式を丸め込むと，式 (1) の Lincoln-Petersen 推定量となる [6]．N の MAP 推定値（表 1）を見ると，POP 数が超幾何分布のモード (x = 10) の値となった場合は，N のMAP推定値が真値に一致していることがわかる（不偏推定量ではない）.

6. ポアソン分布による近似

超幾何分布は，母数が 3 個もあり，多い．確率質量関数も二項係数を含んでいて複雑である．超幾何分布の母数 n と N の比をとって，別の母数 p を定義してみよう．この p (= n/N) は，採集された

式(7)

$$LL(N|x,n,m) = \ln n! + \ln(N-n)! + \ln m! + \ln(N-m)! - \ln x! - \ln N! - \ln(n-x)! - \ln(m-x)! - \ln(N-n-m+x)!$$

$$\therefore \frac{\partial LL}{\partial N} = \frac{\partial}{\partial N}[\ln(N-n)! + \ln(N-m)! - \ln N! - \ln(N-n-m+x)!]$$

$$= \frac{\partial}{\partial N}\left[\ln\left\{\sqrt{2\pi(N-n)}\left(\frac{N-n}{e}\right)^{N-n}\right\} + \ln\left\{\sqrt{2\pi(N-m)}\left(\frac{N-m}{e}\right)^{N-m}\right\} - \ln\left\{\sqrt{2\pi N}\left(\frac{N}{e}\right)^{N}\right\} - \ln\left\{\sqrt{2\pi(N-n-m+x)}\left(\frac{N-n-m+x}{e}\right)^{N-n-m+x}\right\}\right]$$

$$= \frac{\partial}{\partial N}\left[\left(N-n+\frac{1}{2}\right)\ln(N-n) + \left(N-m+\frac{1}{2}\right)\ln(N-m) - \left(N+\frac{1}{2}\right)\ln N - (N-n-m+x)\ln(N-n-m+x) + x\right]$$

$$= \frac{1}{2(N-n)} + \ln(N-n) + \frac{1}{2(N-m)} + \ln(N-m) - \frac{1}{2N} - \ln N - \frac{1}{2(N-n-m+x)} - \ln(N-n-m+x)$$

$$\frac{\partial LL}{\partial N} = 0 \Rightarrow \frac{1}{N-n} + \frac{1}{N-m} - \frac{1}{N} - \frac{1}{N-n-m+x} = 2\ln\frac{N(N-n-m+x)}{(N-n)(N-m)}$$

式(8)

$$\mathrm{HG}(x|N,n,m) = \frac{{}_n\mathrm{C}_x \times {}_{N-n}\mathrm{C}_{m-x}}{{}_N\mathrm{C}_m} = \frac{n^{\underline{x}}/x! \times (N-n)^{\underline{m-x}}/(m-x)!}{N^{\underline{m}}/m!} = {}_m\mathrm{C}_x \frac{n^{\underline{x}}(N-n)^{\underline{m-x}}}{N^{\underline{m}}}$$

$$= {}_m\mathrm{C}_x \frac{(pN)^{\underline{x}}(N-pN)^{\underline{m-x}}}{N^{\underline{m}}} = {}_m\mathrm{C}_x \frac{(pN)^{\underline{x}}(qN)^{\underline{m-x}}}{N^{\underline{m}}}$$

$$= {}_m\mathrm{C}_x \frac{(pN)(pN-1)\cdots(pN-x+1) \times (qN)(qN-1)\cdots(qN-m+x+1)}{N(N-1)\cdots(N-m+1)}$$

$$= {}_m\mathrm{C}_x \frac{p\left(p-\frac{1}{N}\right)\cdots\left(p-\frac{x-1}{N}\right)q\left(q-\frac{1}{N}\right)\cdots\left(q-\frac{m-x-1}{N}\right)}{1\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{m-1}{N}\right)}$$

ここで$N \to \infty$の極限を取る

$$= {}_m\mathrm{C}_x p^x q^{m-x} = {}_m\mathrm{C}_x p^x (1-p)^{m-x} = \mathrm{B}(x|p,m).$$

親個体が全親個体のうちで占める割合を意味する．そして，Nに関する極限を取ると，超幾何分布 HG$(x|N,n,m)$から二項分布 B$(x|p,m)$が導出される（下線は下降階乗冪を意味する）：式 (8)

ここでなされた近似は，たとえばN＝1,000，n＝100の時に見られるxのふるまいと，N＝100,000，n＝10,000 の時に見られる x のふるまいの違いを気にしない（サンプリングエラーの規模の違いを無視する）ということを意味している．これで，母数が 3 個から 2 個になった．

式(9)

$$\mathrm{B}(x|p,m) = {}_m\mathrm{C}_x p^x (1-p)^{m-x} = \frac{m!}{x!\,(m-x)}\left(\frac{\lambda}{m}\right)^x \left(1-\frac{\lambda}{m}\right)^{m-x}$$

$$= \frac{1}{x!}\frac{m(m-1)\cdots(m-x+1)}{m^x}\lambda^x \left(1-\frac{\lambda}{m}\right)^m \Big/ \left(1-\frac{\lambda}{m}\right)^x$$

$$\lim_{m\to\infty}\frac{m(m-1)\cdots(m-x+1)}{m^x} = \lim_{m\to\infty}\left(1\cdot\frac{m-1}{m}\cdot\frac{m-2}{m}\cdots\frac{m-x+1}{m}\right) = 1,$$

$$\lim_{m\to\infty}\left(1-\frac{\lambda}{m}\right)^m = e^{-\lambda},\ \lim_{m\to\infty}\left(1-\frac{\lambda}{m}\right)^x = 1.$$

$$\therefore \lim_{m\to\infty} {}_m\mathrm{C}_x p^x (1-p)^{m-x} = \frac{e^{-\lambda}\lambda^x}{x!} = \mathrm{Pois}(x|\lambda)$$

POP 数(x)	生起確率	N の MAP 推定値	95%CI 下限	95%CI 上限
3	0.008	333333	179302	3768170
4	0.019	250000	138410	1612560
5	0.038	200000	114060	917501
6	0.063	166667	97587	594052
7	0.090	142857	85694	451682
8	0.113	125000	76571	355072
9	0.125	111111	69335	289510
10	0.125	100000	63438	242989
11	0.114	90909	58531	208533
12	0.095	83333	54376	182110
13	0.073	76923	50807	161275
14	0.052	71429	47706	144467
15	0.035	66667	44983	130651
16	0.022	62500	42571	119113
17	0.013	58824	40420	109344
18	0.007	55556	38486	100978
19	0.004	52632	36739	93738
20	0.002	50000	35152	87418
21	0.001	47619	33703	81856
22	0.000	45455	32374	76927
23	0.000	43478	31150	72323
24	0.000	41667	30021	68505

表 2　$\lambda = mn/N$としてポアソン分布から計算されるPOP数 xの生起確率 (N=100,000; n=1,000; m=1,000)と，各POP数 xの下で計算された親の総数 Nの事後分布から求めた 95%信用区間 (n=1,000; m=1,000).

さらにもう一段階，近似を進めてみる．$\lambda = mp$ という母数を新たに定義すると，親の全個体数，採集された親個体数，採集された子個体数という三母数の比が常に一定であるということが仮定されることになる．そして，mに関する極限を取ると，二項分布 $\mathrm{B}(x|p,m)$ からポアソン分布 $\mathrm{Pois}(x|\lambda)$ が導出される：式 (9)

繰り返しになるが，上でなされた近似の意味は，たとえば N = 1,000，n = 100，m = 100 の場合と N = 100,000，n = 10,000，m = 10,000 の場合とで，x のふるまいに違いがないことを仮定するということである．ポアソン分布まで近似を進めることで，母数は 1 個にまで減った．ポアソン分布の母数 λ は，$p = n/N$ なので，$\lambda = mn/N$ である．これは，式 (2) の右辺に一致する．つまり，ポアソン分布の母数は，超幾何分布に従う確率変数（すなわち POP 数 x）の期待値になっているのだ．ポアソン分布で近似した場合の Nの推定値を表 2 に示し

たので，表 1 の結果と比較して，近似の善し悪しを評価していただきたい．

7. 今後の作業

本稿で紹介した計算は，クロスキンデータを用いた資源量推定の「序論の導入部」である．今回計算に用いた超幾何分布は，第 2 節で置いた一子相伝の仮定の下でのみ成り立つ．一個体の母親から生まれて生き残る子の数（繁殖成功）は，実際には何らかの確率分布（たとえばポアソン分布）に従うわけだが，この分布を特定した上で，式 (6) に導入する必要がある．本稿では，親子のペア数（POP）に焦点を当てたが，同時に半同胞のペア数（HSP: Half-Sibling Pair）も考慮することで，推定精度の向上が期待される．

多年生で繰り返し繁殖する生物を想定した場合，親の世代は均一でないことをモデルに組み込まなければならない．親の齢を体長から逆算して決める場合は，その推定の誤差も考慮する必要性がある．また，クロスキンデータを何年間にもわたって取り続けるならば，ある年の N の事後分布を次の年の事前分布にするベイズ更新の方法が利用できる．その場合には，その世代の親が一年間に被った死亡の影響も加味することになるが，これに伴う不確実性もモデルに組み込む必要がある．

第 3 節で紹介した近親判別に伴う誤差もまた，モデルに組み込まれるべきである．既に挙げただけでも，N の区間推定値を広くする誤差の要因は数多く，かつ多様である．こういった多段階にわたる誤差は，階層ベイズモデリングによって対処することで，数値計算の上での取り扱いが楽になるであろう．

クロスキン法による資源量推定値にもっとも深刻なバイアスをもたらすと考えられるのが，生物の移動である．齢や季節によって規則性があるにもかかわらず，移動パターンに関する十分な情報が得られない場合，サンプリングの空間的な偏りが推定値のバイアスに結びつくという懸念は，杞憂でありえない．このバイアスをより直感的に理解するためには，ある個体の親は，ある特定の海域でサンプリングされる確率が非常に高く，のこりの海域にいる確率はほとんどないという状況を想像すればよい．それにもかかわらず，そういった空間分布のムラを無視して，あたかもどの海域でも等しい確率でサンプリングされるかのようなモデルを立てて推定値を計算すると，資源量が過大評価されたり過小評価されてしまうだろう．しかしながら，生物の移動パターンについての情報が不十分であるならば，このバイアスを効果的に取り除く術は，今のところ思い当たらない．

謝辞：鈴木伸明博士をはじめとする太平洋くろまぐろ近親遺伝分析（Close-Kin 分析）関係者のみなさまに深謝いたします．荒木仁志博士，岡村寛博士，酒井一彦博士，森本直子博士，東京大学大気海洋研究所資源解析分野のみなさまには大変貴重なコメントをいただきました．

参考文献

[1] Bravington MV, Simon NJ, Skaug HJ (2014a) Antarctic Blue Whale surveys: augmenting via genetics for close-kin and ordinal age. Report SC65b/SH117 submitted to the Scientific Committee of the International Whaling Commission.

[2] Bravington MV, Grewe PG, Davies CR (2014b) Fishery-independent estimate of spawning biomass of Southern Bluefin Tuna through identification of close-kin using genetic markers. FRDC Report 2007/034. CSIRO, Australia.

[3] Seber GAF (1982) The estimation of animal abundance and related parameters. 2nd ed. The Blackburn Press. Caldwell, NJ.

[4] Gazey WJ, Staley MJ (1986) Population estimation from mark-recapture experiments using a sequential Bayes algorithm. Ecology 67: 941-951.

[5] Yang X, Pal N (2010) Estimation of a population size through capture-mark-recapture method: a comparison of various point and interval estimators. Journal of Statistical Computation and Simulation 80: 335-354.

[6] Seber GAF (2006) Capture-recapture methods - I. In Encyclopedia of Statistical Sciences, Kotz S, Balakrishnan N, Read CB, Vidakovic B, eds. 2: 1-7.