論文の内容の要旨

論文題目　Statistical Methods for Ecological and Epidemiological Data
生態学および疫学データのための統計的方法

氏　　名　小谷野　仁

本論文は、(i) 生物多様性の推定、(ii) 小地域推定、および (iii) 個体群密度の推定という、生態学と疫学における3つの問題を、数理統計学と確率論を用いて扱うための方法論に関する著者の研究の結果をまとめたものである。論文の第1章では、本論文が扱う問題とその背景および先行研究について述べ、論文全体の構成を示す。

第2章と第3章では、近年急速に蓄積されている配列データを用いて生物多様性 (α多様性) を定量するための方法論を提案し、微生物生態学への応用によってその有効性を示し、更に方法論の背後にある数理的な基礎を組織的に述べる。1つの環境下の全てのリボソーム RNA 遺伝子配列を収集することは到底不可能であるから、我々は標本から母集団の多様性を推定するしかないが、この時、(i) 母集団の多様性をどのように定義するかと (ii) それを標本からどのように推定するかという2つの問題がある。伝統的には、豊富さと均等度が、多様性の指標が反映すべき最も重要な性質であった。しかし、配列の全ての組の間のダイバージェンスを考慮して多様性を測定しようとする時には、カテゴリー間の均等度は重要でなくなる。また、配列の群集は一般にいくつかの部分群集からなるため、その多様性を、各部分群集の中の多様性と部分群集の間の多様性の

両方を反映する階層的な量として定義することが望まれる。例えば、母集団からランダムに選んだ配列の間の距離の平均を計算する通常の方法では、配列群集が階層を形成するという側面が考慮されていない。そこで、我々は、配列群集の母集団の多様性を豊富さと階層形成性を反映する量として定義し、それを推定する方法論を提案する。

我々が生物多様性の定量問題に接近する際の基本的な着想は、大雑把には次のようである。$\overline{A}$ を 4 つの文字 a, g, c, および t と空文字 e からなる集合とし、A^* を $\overline{A}$ 上の文字列の全体とする。A^* 上に Levenshtein 距離 d_L を定義し、距離空間を構成する。文字列 s と r の間の Levenshtein 距離とは、s を r に変形するのに必要な削除、挿入、および置換の 3 つの操作の最小回数のことである。A^* は連接によって非可換な半群をなすが、ベクトル空間ではないから、s_1, …, $s_n \in A^*$ に対して、これらの平均は定義されない。そこで、位置の尺度としてコンセンサス配列 ($m(s_1, \cdots, s_n)$によって表す) をとると、散らばりの素朴な尺度を、例えば

$$\frac{1}{n}\sum_{i=1}^{n} d_L(s_i, m(s_1,...,s_n)) \quad \text{または} \quad \frac{1}{n}\sum_{i=1}^{n} d_L(s_i, m(s_1,...,s_n))^2$$

によって定義できる。我々は、A^*上に確率論を展開することによって、リボソーム RNA 遺伝子配列を用いた生物多様性の定量問題に接近する。我々は、1 つの環境下の微生物全体の多様性の定量に我々の方法論を応用し、多様性と様々な環境パラメータの間の関係を調べる。

第 2 章は、我々の方法論の提案とそれの微生物生態学への応用である。我々は、まず第 2.1 節において、様々な環境下で収集された微生物のリボソーム RNA 遺伝子配列の環境標本をグラフによって視覚化し、配列の群集の構造を捉える。この結果に基づいて、第 2.2 節で生物多様性の定量の方法論を提案する。その数理的な側面は第 3 章で組織的に述べられる。第 2.3 節では、生物多様性の定量に必要になる配列の分類アルゴリズムを提案し、その性能を調べ、第 2.4 節では、提案する多様性の推定量の頑健性を数値的に検討する。そうして第 2.5 節で、我々の方法論をいくつかの極限環境と消化器官の微生物群集に適用して、環境パラメータと微生物多様性の間の関係を考察する。更に第 2.6 節では、環境間の配列の共有量の推定を行って、環境と微生物群集の組成の関係を調べる。補足の図が第 2.7

節で示され、分析に用いられた配列データの出典の一覧が第 2.8 節で与えられる。

第 3 章では、生物多様性の測定のために開発した統計理論を体系的に述べる。文字列データの統計に関しては、著者の知る限り、理論的枠組みさえまだ確立されていない。そこで、まず第 3.1 節と第 3.2 節において、確率文字列の統計の枠組みを提案し、次に第 3.3 節において、いくつかの基礎的な補題を証明する。その後第 3.4 節において、確率文字列の列に対する大数の強法則と、我々が提案する生物多様性の推定量に対する漸近的結果を証明する。これらの結果は、生物配列だけでなく、一般の文字列データの統計的分析においても基礎的な役割を果たすと期待される。

第 4 章では、母集団がいくつかの部分母集団に分かれている場合に、部分母集団の、例えば個体数、平均的な体長、平均年齢などを合わせた多次元の特性量を同時に推定する問題を考える。このような推定は小地域推定と呼ばれ、部分母集団は小地域と言われる。小地域統計学における最も基本的な問題は、小地域の標本の大きさが一般にそれほど大きくないために、各小地域に通常の推定量を適用すると、それが決して小さくない標準誤差をもたらすことにある。小地域推定においては、この問題に対処して推定量の精度を高めるために、データを合併したり平滑化したりする方法として、混合線型モデルが使われてきた。1 変量の混合線型モデルは非常に古くから研究されており、また多変量の混合線型モデルに関しても、釣り合い型の場合には、多くの結果が知られている。

しかし、生態学や疫学の研究においては、標本の大きさが小地域によって異なる非釣り合い型の場合が多くある。そこで第 4 章では、我々は、異なる繰り返し数を持つ多変量混合線型モデルにおける予測問題を、統計的決定理論の枠組みで考察し、ミニマックス性に関するいくつかの結果を与える。我々は、まず第 4.1 節において、考察する問題を定式化し、非釣り合い型の場合の取り扱いが釣り合い型の場合とどのような点で技術的に異なるのかを述べ、その後、続く節で取り扱う推定量を定式化する。次に第 4.2 節で、通常の推定量である小地域毎の標本平均と定式化した推定量のリスクの差の評価を導出し、第 4.3 節で、その評価を使って、標本平均を改良する Efron-Morris 型の 2 つの経験 Bayes 推定量を構成する。第 4.4 節では、我々は、構成した 2 つの経験 Bayes 推定量が標本平均をどの程

度改良しているのかを数値実験によって調べる。最後に第 4.5 節で、我々は、提案する推定量を応用して、アフリカにおけるマラリア感染リスクの解析を行う。

1 つの領域におけるある生物の個体群の大きさや密度を評価することは、生態学における基礎的な主題のうちの 1 つであって、様々な生物に対して、その個体群の大きさや密度が算定、あるいは推定されている。この問題に対する方法論上の研究においても、直接的な計数、遠隔探査、統計的推定などの様々な方法が開発されている。ところで、実際の調査においては、領域の広大さや調査の費用のために、観測を行うべき領域全体に渡って調査を行うことが不可能である場合がある。そこで、我々は、第 5 章において、個体群密度の推定問題と観測領域の決定問題を合わせて考察して、それらの両方を同時に最適化する方法を提案し、その方法の理論的な基礎を与える。我々は、問題を Poisson 過程の平均インテンシティーの逐次推定問題の変形として定式化するが、個体群密度の推定においては、観測領域の範囲は、ある場合には熱帯雨林や海洋の 1 つの領域であり、また別の場合には少量の土壌や海水であるなど、地理的な距離や長さの意味を持ち、その測定の単位は場合によって様々である。そこで、我々は、観測領域の決定規則と個体群密度の推定量に対して、スケール変換群の下での共変性を要請する。不変性原理を用いる結果、我々は、提案する方式の最適性を、漸近的方法を用いる逐次推定の従来の枠組みでではなく、統計的決定理論の枠組みで議論する。

我々は、まず第 5.1 節において、考察する問題を定式化し、その後、時間パラメータ空間上のスケール変換群の下での逐次方式の共変性を定義する。次に第 5.2 節で、Poisson 過程の平均インテンシティーに対する、スケール変換群の下で共変な逐次方式を構成する。この逐次方式の構成の際に使われるいくつかの基礎的な関係式は、第 5.4 節で与えられる。第 5.3 節では、我々は、まず提案する逐次方式の許容性に関する結果を、情報量不等式を用いる方法によって証明し、その後、その逐次方式は、スケール変換群の下で不変な事前分布に対する Bayes 逐次方式であることを示して、ミニマックス性に関する結果を与える。