統計学におけるモデル

—情報量基準の観点から—

山口健太郎*

Abstract

Within the framework of statistics, the goodness of statistical models is evaluated by criteria for model selection, such as the Akaike and Bayesian information criteria. Each information criteria is based on likelihoodist's or Bayesian conception. Here, I analyse the inferences used in the derivation of these criteria, and argue that the goodness, evaluated by the Akaike or Bayesian information criteria reflects frequentist's conception, which is not explained by likelihoodist or Bayesian.

1. 序

コイン投げのように，確率的な振る舞いをする現象をモデル化すること，すなわち現象を記述する確率分布族を特定 (specify) し，その母数を推定するのが，統計学の仕事の 1 つである．しかし，ある確率的な現象から異なる統計的モデルが導出されたときに，どのモデルが（統計学の観点から）「よい」モデルであるということができるのだろうか．従来の統計学の観点から言えば，たとえばカイ 2 乗適合度検定のように，確率的な現象から得られるデータにもっとも合致するモデルこそが，よいモデルであった．しかし，パラメータの数が多い確率分布族をモデルとして想定すれば，データにより合致させることが可能なため，モデルが複雑過ぎるにもかかわらず，「よい」モデルとして選択されてしまうことになる．

赤池が提示した赤池情報量基準 (AIC) は，従来とは異なるモデルのよさを判定する基準である．すなわち，データとモデルとの適合度だけでなく，パラメータ数もモデルのよさを判定する基準に入っているので，パラメータ数が比較的多くならず，なおかつデータと合致するようなモデルを選択するような基準になっている．

本稿では，AIC のようなモデル選択基準の要請するモデルの「よさ」が，どういったものであるかについて論じる．論文の構成は以下のとおりである．2 章では，統計的モデルとはどのようなもので，統計学の枠組みの中でモデル選択の問題がどのように定式化されるのかについて，簡単に説明する．3 章では，AIC がどのような推論に基づいて導出されるのかを再構成することで，どういった「よさ」を判定する基準になっているかについて考察する．4 章では，ベイズ統計学の考え方に基づいて導出されたモデル選択基準である BIC について論じる．5 章では，

* 京都大学大学院文学研究科博士課程 Kentaro.Yamguchi@l01.mbox.media.kyoto-u.ac.jp

AIC と BIC とを比較検討することで，双方の情報量基準が共通の困難を抱えていることを示したい．

2. 統計学におけるモデルとその選択

最初に，統計学とはどのような方法論で，そこで登場するモデルとはどのようなものなのかについて説明したい．統計学とはデータを処理するための方法論である．どのようなデータを処理するのかと言えば，たとえば，人の身長やサイコロを投げたときの出る目や日経平均株価などのように，数値化されるデータが，統計学で取り扱われる対象である．

このようなデータを大量に集めて，その集団的な性質を調べるのが統計学である．コインの例を考えてみよう．偏りのないコインを 10000 回投げて，表が出るか，裏が出るかを記録していくとする．このとき，コインの表が出るか，それとも裏が出るかどうかは，でたらめに見える．たとえば，表表裏裏裏表裏裏表裏$\cdots$ のようになる．しかし，10000 回投げたときに，表が出る頻度は，ほぼ $\frac{1}{2}$ になるであろう．このように，データの発生が確率的に変動する場合に，その背後にある集団的な規則性を確率を用いて記述あるいは推測する方法論が，統計学である．

上と同じ例を使って，確率的な (stochastic) 振る舞いをする現象がどのように確率を用いて記述されるのかをみる[*1]．表が出る確率を p とすると，裏が出る確率は $1-p$ となる[*2]．このようにおくと，コインを n 回投げたときに表が k 回出る確率を求めることが可能である．というのも，コインの表が出る回数が二項分布になることが知られているからである．すなわち，i 回めに表が出た場合，確率変数を $X_i = 1$ とし，裏が出た場合，確率変数を $X_i = 0$ とする．すると，表が出る回数は，確率変数 $X = X_1 + \cdots + X_n$ によって表すことができる．このとき，確率変数の値ごとに，それが実現する（この場合，表が確率変数 X の値だけ出る）確率が決まっている．つまり，X の分布は二項分布にしたがっていると知られているので，その確率は，

$$p(X = k) = \binom{n}{k} p^k (1-p)^k \tag{1}$$

のように表すことができる．もしコインの表の出る確率が知られているならば，n 回投げたときに表が k 回出る確率を求めることができる．たとえば，コインの表の確率が $\frac{1}{2}$ であると知られている場合，10 回投げたときに表が 6 回出る確率は，

$$\begin{aligned} p(X = 6) &= \binom{10}{6} \frac{1}{2}^6 \times \left(1 - \frac{1}{2}\right)^4 \\ &= \frac{210}{1024} \end{aligned}$$

となる．以上のようにして，確率的な振る舞いを，確率を用いて記述できる．

[*1] 本来ならば，stochastic と probability とを訳し分けなければいけないが，ともに，「確率」という定訳がついているので，本稿もそれにしたがいたい．

[*2] 本来，統計学について検討する場合，それは確率の解釈の問題と切り離せない．しかし，ここで用いられている確率は，統計的確率と呼ばれるもので，確率の解釈に依存せずに，考えられる確率である [内井 1995]．なお，確率についての数学的な説明は省略するので，詳しくは，たとえば [竹村 1991] を参照されたい．

ここで重要な前提は次の 2 つである．すなわち，

(i) コインの出る確率が二項分布にしたがっていること

(ii) コインの出る確率が $\frac{1}{2}$ であること

このように仮定をおいたならば，n 回コインを投げたときに，k 回表が出る確率を求めることができる．

(i) のように，想定された特定の確率分布族のことを（統計的）**モデル族**と呼び，その要素である各確率分布を**統計的モデル**(statistical model) と呼ぶ．また，分布族において個々の分布を指定するものを**パラメータ**と言う．コインの例だと，$\{Bi(n,p)\}_{n\in\mathbb{N},p\in[0,1]}$ がモデル族，n,p がパラメータであり，n,p が指定された $Bi\Big(10000,\frac{1}{2}\Big)$ がモデルとなることに注意されたい[*3]．また，統計学では，真であるモデルを想定することがあり，確率分布とパラメータとが真であるモデルを**真のモデル**と呼ぶことにする．

いま，(i), (ii) の仮定がおけない場合に，あるコインによるコイン投げに関するモデルを特定する問題を考える．(i) が既知であることを仮定できるならば，パラメータ p の値が未知であったとしても（すなわち (ii) のような仮定をおくことができない場合），試行から得られるデータを集めることで，パラメータを推定することができる．あるいは，仮説検定，すなわち帰無仮説 $H:p=13$ をたて，試行から得られるデータから，この帰無仮説が棄却されるか採択されるかどうかを検定にかければ，モデルが得られる．コイン投げのような簡単な状況では，モデル族の特定は容易であるので，結果としてモデルを推定することも簡単である．しかし，データの発生が複雑な状況（たとえば，日経平均株価の変動など）の場合，モデル族の特定は難しい．仮に，発見的手法でモデル族の特定を行えたとしても，それよりもよいモデル族が存在するかもしれない．これにともなって，モデルの特定も困難になる．

このようなデータの発生が複雑な状況の場合には，モデルを特定する方法として次のようなものを考える．まず，現象を記述する複数のモデル族を（発見的手法であれ）用意する．そして，モデル族を代表するモデルを選び出し（具体的には，あとで見るように，モデル族のもつパラメータを推定し，その推定したパラメータをもつモデルがモデル族を代表するモデルとなる），それらからもっとも「よい」モデルを選択することである．これが，統計学における**モデル選択**の問題であり，実際，次章以降で論じる AIC や BIC において想定されているモデル選択はいま述べたものになっている．このような問題設定は，ほかのモデル選択の問題設定の可能性を否定するものではない[*4]．

[*3] 統計学に限って言えば，確率分布族のことを統計的モデルと呼ぶ場合 [竹村 1991] と，その要素である確率分布を統計的モデルと呼ぶ場合 [小西 2004] の 2 つの流儀がある．本論文では，[小西 2004] の流儀を採用した．どちらの流儀を採用しても議論の本質が変わらないが，情報量基準にしたがって最終的に選択されるのがある確率分布であり，それをモデルと呼ぶほうが議論を進めやすいというのが，その理由である．

[*4] つまり，本稿で想定している統計学におけるモデル選択の問題は，ある現象を記述する複数のモデル族が用意されているということが前提されていて，このようなモデル族がどのように用意されるのかについての指針は与えられていない．

3. 赤池情報量基準 (AIC)

では，どのような方法によって，複数のモデル族からもっとも「よい」モデルが選択できるのか．そのためには，もっともよいモデルを選択するための基準を導入する必要がある．これを実現するために，モデルのよさを何らかの形で数値化し，その数値の比較によってモデルのよさを判定する．このようなモデルのよさを判定するための基準の 1 つが**情報量基準**であり，その中に以下の章で取り上げる AIC や BIC がある．本章では，赤池が提唱した AIC を取り上げるが，それを導出するために，まずカルバック–ライブラー情報量を導入したい．

3.1 カルバック–ライブラー情報量

まず出発点として，真である確率分布との距離の大小によって，モデルのよさを測ることを考える．このような距離を測る測度の 1 つとして**カルバック–ライブラー情報量**(Kullbuck-Leibler information) を導入する．カルバック–ライブラー情報量は次のように定義される[*5]．

定義 1. 確率密度関数 $g(x)$ を真のモデルとし，確率密度関数 $f(x)$ をわれわれが想定したモデルとする．このとき，カルバック–ライブラー情報量 $I(g;f)$ は，

$$I(g;f) \stackrel{\text{def}}{=} E_g\left[\log\left\{\frac{g(X)}{f(X)}\right\}\right] \tag{2}$$

となる．ここで，E_g は確率分布 g に関する期待値をあらわす[*6]．

確率変数が連続の場合，

$$I(g;f) = \int_{-\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx \tag{3}$$

となる．この $I(g;f)$ によってモデルのよさを測る．すなわち，真のモデル $g(x)$ を仮定すると，$I(g;f)$ を小さくするモデル $f(x)$ が「よい」モデルとなる．

真のモデルが既知の場合には，$I(g;f)$ を求めることが可能である．しかし，真のモデルが未知の場合には，$I(g;f)$ を求めることができない．すなわち，式 (3) より，

$$\begin{aligned} I(g;f) &= E_g[\log g(X)f(X)] \\ &= E_g[\log g(X)] - E_g[\log f(X)] \end{aligned} \tag{4}$$

[*5] カルバック–ライブラー情報量によって，なぜ f, g のあいだの距離を測ることができるのかに関しての説明は省略する．詳しくは，[小西 2004], [下平 2004] を参照されたい．

[*6] 確率変数 X の期待値は，

$$E[X] \stackrel{\text{def}}{=} \begin{cases} \sum_x xf(x) & X \text{ が離散の場合} \\ \int xf(x)dx & X \text{ が連続の場合} \end{cases}$$

のように定義される．以下，確率分布関数の期待値をあらわす場合には，E_f のように，確率密度関数で添字づけする．

となるが，g が未知であるならば，このような期待値を求めることができない．

いま，式 (4) の第 2 項 $E_g[\log f(X)]$ だけに絞って議論を行いたい．というのも，第 1 項は，想定されるモデルに依存しないので，どのようなモデルを想定しても同じ値となるからである．このことは，第 1 項 $E_g[\log g(X)]$ を $I(g;f)$ から差し引いたとしても，モデルの「よさ」についての順序関係に変化がないことを意味する．以下，第 2 項だけに絞って考察する．

第 2 項をみると，真の分布 g が既知でないと，値は求められないことがわかる．というのも，

$$E_g[\log f(X)] = \int_{-\infty}^{\infty} g(x) \log f(x) dx \tag{5}$$

のように，真の密度関数 $g(x)$ に依存しているためである．このため，$E_g[\log f(X)]$ を近似的に求めることを考える．そこで，データ $x_1, \ldots, x_n$ の発生が等確率である確率分布 $\widehat{g}$（**経験分布**と呼ばれる）を仮定すると，

$$E_{\widehat{g}}[\log f(X)] = \frac{1}{n} \sum_{\alpha=1}^{n} \log f(x_\alpha) \tag{6}$$

となり，想定されるモデル f とデータ $x_1, \ldots, x_n$ だけで，$E_g[\log f(X)]$ の近似的な値を求めることができる．問題はこの $E_{\widehat{g}}[\log f(X)]$ が $E_g[\log f(X)]$ の近似になっているかどうかだが，それを保証するのが確率収束という概念である．

定義 2. 確率変数の列 $(X_n)_{n\in\mathbb{N}}$ が確率変数 X に**確率収束**するとは，

$$\forall \varepsilon > 0 \left(\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0 \right) \tag{7}$$

となることを言う．

確率収束という概念を導入すると，確率変数 $\log f(X_n)$ の確率分布 $\widehat{g}$ に関する期待値が $E_g[\log f(X)]$ に確率収束することから，$E_{\widehat{g}}[\log f(X)]$ が $E_g[\log f(X)]$ の近似になっていることが保証される．したがって，カルバック–ライブラー情報量をあらわす式 (4) は次のように近似されたことになる．

$$I(g;f) \approx E_g[\log g(X)] - \frac{1}{n} \sum_{\alpha=1}^{n} \log f(x_\alpha) \tag{8}$$

本節の議論をまとめると，以下のようになる．カルバック–ライブラー情報量 $I(g;f)$ を，真の分布 g と想定されるモデル f とのあいだの近さを測る指標として導入したが，真のモデル g は未知であるので，$I(g;f)_{approx}$ とでも言うべき量（すなわち，式 (8) の右辺）しか求められず，これがモデルの「よさ」を測る基準となっている，ということである．

3.2 赤池情報量基準

前節で与えた $I(g;f)_{approx}$ は，真のモデル g や想定されたモデル f に関する情報をまったく用いずに求められたものである．以下では，モデルの具体的な形にもう少しコミットすることで，$I(g;f)_{approx}$ がどのように表されるかについて見ていきたいと思う．

いま，想定されているモデル族を $\{f(x|\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\subset\mathbb{R}^p}$ とすると，n 回の試行から得られるデータ $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ からパラメータ $\boldsymbol{\theta} = {}^t(\theta_1, \ldots, \theta_p)$ を推定することができる．パラメータ $\boldsymbol{\theta}$ の推定方法はいろいろ存在するが，赤池が採用する方法は**最尤推定法**である．すなわち，尤度が最大となるときのパラメータの値を，推定値として決定する方法である．では，尤度とは何か．**尤度**(likelihood) とは，データを固定し，パラメータ $\boldsymbol{\theta}$ の関数とみなした

$$l(\boldsymbol{\theta}) \stackrel{\text{def}}{=} f(\boldsymbol{x}_n|\boldsymbol{\theta}) \tag{9}$$

のことを言う[*7]．ここで，尤度関数 l が試行によって集められたデータに依存することに注意されたい．もう少し詳しく言うと，いま確率的な振る舞いをする現象を取り扱っているので，試行が違えば，それによって集められるデータも違う．そのため，尤度関数が変化する．また，集められるデータ数も異なれば，それに伴って尤度関数が変化することにも注意されたい．

このように尤度関数を定義すると，最大尤度は，

$$l(\widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l(\boldsymbol{\theta}) \tag{10}$$

によって与えられる．この最大尤度を与える $\widehat{\boldsymbol{\theta}}$ が，パラメータ $\boldsymbol{\theta}$ の推定量となる[*8]．ここで，真のモデルとして $g(z)$，想定されるモデルとして $f(z|\widehat{\boldsymbol{\theta}})$ を考えると，前節より，真のモデルと想定されるモデルとの近さは，

$$I(g(z); f(z|\widehat{\boldsymbol{\theta}})) \approx E_g[\log g(Z)] - \frac{1}{n}\sum_{\alpha=1}^{n} \log f(x_\alpha|\widehat{\boldsymbol{\theta}}) \tag{11}$$

によって測ることができる．いま，式 (11) を上で定義した最尤推定量 $\widehat{\boldsymbol{\theta}}$ を用いてあらわすことを考える．式 (11) の第 2 項は $\dfrac{1}{n}\displaystyle\sum_{\alpha=1}^{n} \log f(x_\alpha|\widehat{\boldsymbol{\theta}})$ だが，これが最大対数尤度によってあらわされることに注意されたい．すなわち，最大尤度は定義から，

$$l(\widehat{\boldsymbol{\theta}}) = \prod_{\alpha=1}^{n} f(x_\alpha|\widehat{\boldsymbol{\theta}}) \tag{12}$$

となるので，最大対数尤度は，$\displaystyle\sum_{\alpha=1}^{n} \log f(x_\alpha|\widehat{\boldsymbol{\theta}})$ となる．よって，$E_{\hat{g}}[\log f(Z|\boldsymbol{\theta})] = \dfrac{l(\widehat{\boldsymbol{\theta}})}{n}$ となる．以上から，式 (11) の第 2 項 $E_g[\log f(Z|\boldsymbol{\theta})]$ の推定量が，$\dfrac{l(\widehat{\boldsymbol{\theta}})}{n}$ となることが示される．

いま平均対数尤度を次のように定義する．

$$\eta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} E_g[\log f(Z|\boldsymbol{\theta})] \tag{13}$$

ここで定義した平均対数尤度が，カルバック–ライブラー情報量 $I(g; f)$ の第 2 項に対応していることに注意されたい．すなわち，$I(g; f)$ の「よさ」に関する順序が平均対数尤度の大小関

[*7] 尤度関数とも呼ばれるので，以後断りなく，尤度と尤度関数とを区別せず用いる．

[*8] 尤度関数によって推定される $\widehat{\boldsymbol{\theta}}$ もまた集められたデータに依存するので，このことを明示的にするために $\widehat{\boldsymbol{\theta}}(\boldsymbol{X}_n)$ ($\boldsymbol{X}_n$はデータの発生についての確率変数) あるいは $(\widehat{\boldsymbol{\theta}})_n$ と書くこともある．このとき，$(\widehat{\boldsymbol{\theta}})_{n\in\mathbb{N}}$ が確率変数の列になっていることに注意されたい．

係によってあらわされていることになる．しかし，$I(g;f)$ を近似した $I(g;f)_{approx}$ の第 2 項は $\dfrac{l(\widehat{\boldsymbol{\theta}})}{n}$ のように，対数尤度によってあらわされている．

ここで重要なのは，平均対数尤度と対数尤度とのあいだで，「よさ」に関する順序が逆転することである．すなわち，真のパラメータを $\boldsymbol{\theta}_0 = {}^t(\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ とすると，平均対数尤度に基づけば，$\eta(\widehat{\boldsymbol{\theta}}) \leq \eta(\boldsymbol{\theta}_0)$ となるはずだが，対数尤度に基づけば，$\widehat{\boldsymbol{\theta}}$ が最も生じやすいので，$l(\widehat{\boldsymbol{\theta}}) \geq l(\boldsymbol{\theta}_0)$ となる．

以上をまとめると，次のようになる．いま，モデル選択の基準としてカルバック–ライブラー情報量を想定すると，平均対数尤度の大小によって，モデルの善し悪しが判定される．しかし，この平均対数尤度自体は，真の分布 g が未知であることから，求めることができない．そのため，g の代わりに経験分布 $\widehat{g}$ を用い，さらにパラメータとして最尤法から求めた $\widehat{\boldsymbol{\theta}}$ によって，平均対数尤度を近似的に求めることになる．だが，$\widehat{g}$ から求められる $\dfrac{l(\widehat{\boldsymbol{\theta}})}{n}$ の「よさ」に関する順序関係が，平均対数尤度によるよさに関する順序関係と一致しないために，最大尤度だけからでは公正なモデル選択を行うことができないのである（図 1[*9]）．

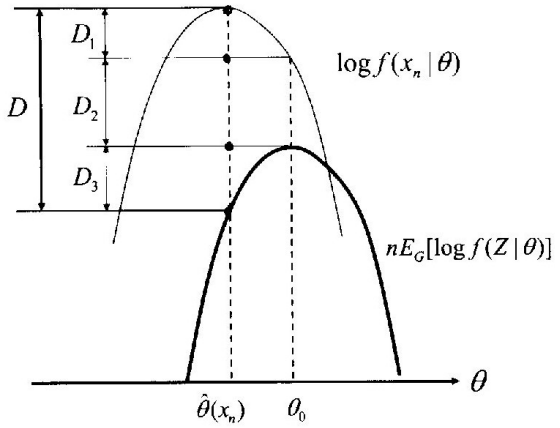

図 1　対数尤度（細線）と平均対数尤度（太線）

このようなことから，対数尤度に基づいて求められる $\log f(Z|\widehat{\boldsymbol{\theta}})$ と，平均対数尤度に基づいて求められる $nE_g[\log f(Z|\widehat{\boldsymbol{\theta}})]$ とのあいだのバイアスを補正しないといけない．バイアスを以下のように定義する．

$$b(g) \stackrel{\text{def}}{=} E_{g(\boldsymbol{x}_n)}[\log f(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{g(z)}[\log f(Z|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}_n))]] \tag{14}$$

ただし，期待値 $E_{g(\boldsymbol{x}_n)}$ は標本分布 $\boldsymbol{X}_n = \{X_1 \ldots, X_n\}$ の同時密度関数 $\displaystyle\prod_{\alpha=1}^{n} g(x_\alpha) = g(\boldsymbol{x}_n)$ に関する期待値であり，期待値 $E_{g(z)}$ は真のモデル $g(z)$ に関する期待値である．図 1 で言うならば，$\log f(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{g(z)}[\log f(Z|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}_n))]$ が，D に相当していることになる．ここで，

[*9] [小西 2004]p.51

$g(\boldsymbol{x}_n)$ に関する期待値をとるのは，各データ $x_1, \ldots, x_n$ はそれぞれ確率変数 $X_1, \ldots, X_n$ の実現値であるが，実現値によってばらつきが生じてしまうので，それを平均化するためである．

このバイアス $b(g)$ を推定によって求め，その分を対数尤度から補正することによって，次のような情報量基準

$$IC(\boldsymbol{X}_n; \widehat{g}) := -2(\text{統計モデルの対数尤度} - \text{バイアスの推定量}) \tag{15}$$

$$= -2 \sum_{\alpha=1}^{n} \log f(X_\alpha|\widehat{\theta}) + 2\{b(g) \text{ の推定量 }\} \tag{16}$$

が構成される．これは，$I(g; f)_{approx}$ の第 2 項の推定量に -2 を乗じたものに相当するのだが，第 1 項が省略されるのは，前節で述べたように，第 1 項が真のモデル g だけに依存しているので，この項を無視しても，モデル選択基準として十分であることによる．いま，$b(g)$ は図 1 のように 3 つの項 D_1, D_2, D_3 に分解できる．

$$\begin{aligned} b(g) = & \underbrace{E_{g(\boldsymbol{x}_n)}[\log f(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - \log f(\boldsymbol{X}_n|\boldsymbol{\theta}_0)]}_{D_1} \\ & + \underbrace{E_{g(\boldsymbol{x}_n)}[\log f(\boldsymbol{X}_n|\boldsymbol{\theta}_0) - \log f(Z|\boldsymbol{\theta}_0)]}_{D_2} \\ & + \underbrace{E_{g(\boldsymbol{x}_n)}[nE_{g(Z)}[\log f(Z|\boldsymbol{\theta}_0)] - nE_{g(Z)}[\log f(Z|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}_n))]]}_{D_3} \end{aligned} \tag{17}$$

このように 3 つの項に分解し，各項の推定量を求めるのだが，結論だけ先に述べると，$b(g)$ は漸近的に p となる．すなわち，$D_2 = 0$ となり，D_3 は漸近的に $\frac{p}{2}$ となること，D_1 は漸近的に $\frac{p}{2}$ となることとから，バイアス $b(g)$ は漸近的に p となる．

ここで，漸近的に $b(g)$ が p になるといったが，これはどういうことか．その前に，分布収束について説明する．

定義 3. 確率変数 Z_n の分布 $F_n(x)$ が特定の連続分布 $F(x)$ に**分布収束**するとは，任意の x について，

$$\lim_{n \to \infty} F_n(x) = F(x) \tag{18}$$

となることを言う．

前節で説明した確率収束とここで説明した分布収束とによって，漸近的の意味を与えることができる．

(i) 最尤推定量 $(\widehat{\boldsymbol{\theta}})_n$ は，$n \to +\infty$ のとき，$\boldsymbol{\theta}_0$ に確率収束する．

(ii) 最尤推定量 $(\widehat{\boldsymbol{\theta}})_n$ に対して，$\sqrt{n}((\widehat{\boldsymbol{\theta}})_n - \boldsymbol{\theta}_0)$ の分布は，$n \to +\infty$ のとき平均ベクトル $\mathbf{0} = (0, \ldots, 0)$，分散共分散行列 $J^{-1}(\boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)J^{-1}(\boldsymbol{\theta}_0)$ の p 次元正規分布に分布収束する．ただし，行列 $I(\boldsymbol{\theta}_0), J(\boldsymbol{\theta}_0)$ は，次式で与えられる要素からなる $p \times p$ 行列の $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

での値とする．

$$I_{ij}(\boldsymbol{\theta}) = \int g(x) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j} dx \tag{19}$$

$$J_{ij}(\boldsymbol{\theta}) = -\int g(x) \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} dx \quad (i,j = 1,\ldots,p) \tag{20}$$

上で示した漸近性を利用し，D_3 が漸近的に $\frac{p}{2}$ になることについて説明したいと思う[*10]．

平均対数尤度の定義（式 (13)）から $\eta(\widehat{\boldsymbol{\theta}}) \equiv E_{g(z)}[\log f(Z|\widehat{\boldsymbol{\theta}})]$ とおいて，$\eta(\widehat{\boldsymbol{\theta}})$ を $\boldsymbol{\theta}_0$ のまわりでテイラー展開すると，

$$\begin{aligned}\eta(\widehat{\boldsymbol{\theta}}) = \eta(\boldsymbol{\theta}_0) &+ \sum_{i=1}^{p} (\widehat{\theta}_i - \theta_i^{(0)}) \frac{\partial \eta(\boldsymbol{\theta}_0)}{\partial \theta_i} \\ &+ \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (\widehat{\theta}_i - \theta_i^{(0)})(\widehat{\theta}_j - \theta_j^{(0)}) \frac{\partial^2 \eta(\boldsymbol{\theta}_0)}{\partial \theta_i \partial \theta_j} + \cdots\end{aligned} \tag{21}$$

となる．図 1 からわかるように，平均対数尤度 $nE_g[\log f(Z|\boldsymbol{\theta})]$ が最大となるのは，$\boldsymbol{\theta}$ が $\boldsymbol{\theta}_0$ のとき，すなわち $\frac{\partial \eta(\boldsymbol{\theta}_0)}{\partial \theta_i} = 0$ の場合であるので，式 (21) は（右辺第 4 項以下を無視したという意味で）近似的に

$$\eta(\widehat{\boldsymbol{\theta}}) \approx \eta(\boldsymbol{\theta}_0) - \frac{1}{2}\, {}^t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{22}$$

と表される．ここで，同時密度関数 $g(\boldsymbol{x}_n)$ の定義から $E_{g(\boldsymbol{x}_n)}\left[\sum_{\alpha=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_\alpha|\boldsymbol{\theta})\right] = nE_{g(z)}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Z|\boldsymbol{\theta})\right]$ より，D_3 は $\eta(\boldsymbol{\theta}_0) - \eta(\widehat{\boldsymbol{\theta}})$ の $g(\boldsymbol{x}_n)$ に関する期待値となるので，D_3 は近似的に

$$\begin{aligned} D_3 &\approx E_{g(\boldsymbol{x}_n)}[nE_{g(Z)}[\log f(Z|\boldsymbol{\theta}_0)] - nE_{g(Z)}[\log f(Z|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}_n))]] \\ &= \frac{n}{2} E_{g(\boldsymbol{x}_n)}[\, {}^t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)] \\ &= \frac{n}{2} E_{g(\boldsymbol{x}_n)}[\mathrm{tr}\{J(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\, {}^t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\}] \\ &= \frac{n}{2} \{J(\boldsymbol{\theta}_0) E_{g(\boldsymbol{x}_n)}[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\, {}^t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)]\} \end{aligned} \tag{23}$$

となる．ここで，最尤推定量 $\widehat{\boldsymbol{\theta}}$ が平均ベクトル $\mathbf{0}$，分散共分散行列 $J^{-1}(\boldsymbol{\theta}_0) I(\boldsymbol{\theta}_0) J^{-1}(\boldsymbol{\theta}_0)$ の p 次元正規分布に分布収束することを利用すると，

$$E_{g(\boldsymbol{x}_n)}[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\, {}^t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)] = \frac{1}{n} J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \tag{24}$$

となるので，

$$D_3 \approx \frac{1}{2} \mathrm{tr}\{I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}\} \tag{25}$$

[*10] D_1 が漸近的に $\frac{p}{2}$ になることもほぼ同じ推論のかたちになっている．ここで，重要なのは，漸近性の意味を詳しく見ることである．

となることがわかる．最後に，$I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)^{-1}$ であることから，$D_3 \approx \dfrac{p}{2}$ が示されたことになる．

重要なのは，D_3 が 2 つの意味で漸近的だということだ．1 つめは，

$$\eta(\boldsymbol{\theta}_0) - \eta(\widehat{\boldsymbol{\theta}}) \approx \frac{1}{2}\, {}^t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{26}$$

であり，これは，$\eta(\boldsymbol{\theta})$ を $\boldsymbol{\theta}_0$ のまわりでテイラー展開したときの，第 4 項以降を無視できるという意味で，漸近的である．もう 1 つは，最尤推定量 $\widehat{\boldsymbol{\theta}}$ が漸近的に p 次元正規分布に分布収束しているという意味で，漸近的である．これらの意味で，$D_3 \approx \dfrac{p}{2}$ というのは，標本数 n が「無限に大きい」場合に成り立つ式であり，そうでない場合は保証されない，ということである．

いずれにせよ，最終的に導出された「よさ」の基準は，

$$IC(\boldsymbol{X}_n; \widehat{g}) = -2 \sum_{\alpha=1}^{n} \log f(X_\alpha | \widehat{\boldsymbol{\theta}}) + 2p \tag{27}$$

である．この $IC(\boldsymbol{X}_n; \widehat{g})$ が，**赤池情報量基準**と呼ばれる．

カルバック–ライブラー情報量から AIC の導出に関してまとめておくと，以下のようになる．まず，カルバック–ライブラー情報量の第 2 項 $E_g[\log f(Z|\boldsymbol{\theta})]$ の推定量として，$\dfrac{l(\widehat{\boldsymbol{\theta}})}{n}$ を用いるのに問題があった．そのため，バイアス $b(g)$ を新たに推定しなければならなかった．このとき，$b(g)$ が漸近的に p と等しくなることから，p が $b(g)$ の推定量になる．この推定されたバイアスを考慮に入れたモデルのよさの基準が AIC になる，ということだ．

3.3 AIC における推論

AIC を導出する際に，どのような推論が行われているかを検討したいと思う．簡単にまとめると，以下のようになる．

(A1) 真のモデル g と想定されるモデル f との近さを測る基準として，カルバック–ライブラー情報量を導入する．

(A2) カルバック–ライブラー情報量の第 2 項だけが問題となるが，真の分布 g は未知であるので，これを近似的に求める．

(A3) 最尤法によって推定した $\widehat{\boldsymbol{\theta}}$ でもって，$E_g[\log f(Z|\boldsymbol{\theta})]$ を推定するのだが，これだとバイアスが生じる．

(A4) バイアスを推定し，その分を補正すると，AIC が導出される．

重要なのは，カルバック–ライブラー情報量に基づいてモデル選択を行えるのは，真の分布 g が既知であることを仮定できる場合であり，そのときには，モデルの善し悪しが平均対数尤度 $E_g[\log f(Z|\widehat{\boldsymbol{\theta}})]$ によって判定されるのだが，真の分布が既知であるという仮定自体をおくことができない，という点である．すなわち，この仮定を前提しなければ，カルバック–ライブラー情

報量によってモデル選択を行うことができないということである．そこで，(A2) の段階で，経験分布関数 $\hat{g}$ を導入したが，この (A2) の過程を経由してしか，真の分布 g と想定されるモデル f との近さを測ることができない，つまりバイアスが生じてしまうので，最尤法だけで真の分布 g と想定されるモデル f との近さを正確に測ることができないのである．そこで，バイアス $b(g)$ を推定するのだが，$E_{\hat{g}}[\log f(Z|\boldsymbol{\theta})]$ は最尤法に基づいて導出される量であるのに対し，バイアス $b(g)$ は，（前に述べたように 2 つの意味で）漸近的に p に等しいということである．しかしこの (A4) でのバイアス $b(g)$ の推定は，最尤法に基づいて保証される推論ではないのである．

これが意味することは，AIC の導出でおこなわれる 2 つの推定が，それぞれ別の基準に基づいたものであり，導出された AIC は統一的な「よさ」の基準にしたがって求められたモデル選択基準を与えているのでは決してない，ということである．つまり，パラメータ $\boldsymbol{\theta}$ は尤度だけに基づいて推定されるのだが，$b(g)$ は漸近性という頻度論的な考え方を「こっそり」もち込んでいるのである．この $b(g)$ の推定量 p の「よさ」は，尤度が最大であることによって保証されているのでは決してなく，標本数 n が「無限に大きい」場合（この部分が頻度論的な保証に相当する）には，尤度が最大であることによって保証されるだけである．しかし，標本数 n は有限個なので，推定量 p の「よさ」を尤度だけから説明されているのではない．にもかかわらず，この部分がどのように尤度主義の立場から正当化できるのかについては，必ずしも明らかでないのである[*11]．

4. BIC

この章では，モデル選択のための別の情報量基準である BIC を取り上げる．BIC とは簡単に言えば，ベイズ統計学に基づいたモデル選択基準であり，AIC と同様にモデルのもつパラメータ数を考慮したものになっている．

4.1 BIC の導出

BIC を導出するための出発点は，ベイズの定理である．いま，r 個のモデル族を $M_1, \ldots, M_r$ とし，i 番目のモデル族が生起する事前確率を $P(M_i)$ とすると，i 番目のモデル族の事後確率は，ベイズの定理より

$$P(M_i|\boldsymbol{x}_n) = \frac{p_i(\boldsymbol{x}_n)P(M_i)}{\sum_{j=1}^{r} p_j(\boldsymbol{x}_n)P(M_j)} \quad (i = 1, \ldots, r) \tag{28}$$

で与えられる．ただし，各モデル族 M_i は確率密度関数 $f_i(x|\boldsymbol{\theta}_i)$ $(\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i \subset \mathbb{R}^{p_i})$ とパラメータ $\boldsymbol{\theta}_i$ の事前分布 $\pi_i(\boldsymbol{\theta}_i)$ によって特徴付けられているとする．このとき，データ $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$

[*11] もっとも，赤池自身は尤度に基づいた推定を「よい」推定であると考えているかと言えば，そうでないように思われる．尤度に基づいた推論を保証するのが尤度原理と呼ばれるものだが，赤池はこの基礎が数学的に保証されたものでないと考えている．([Akaike 1982]) なお，尤度原理についての説明は，たとえば [Edwards 1992], [Hacking 1965] を参照されたい．

に関するモデル族 M_i の周辺尤度

$$p_i(\boldsymbol{x}_n) = \int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \tag{29}$$

は i 番目のモデル族からデータが得られる確からしさと考えられる[*12].

ベイズ統計学の考え方に基づいて，事後確率が最大となるモデル族を採用することにする．いま，事前確率 $P(M_i)$ はすべて等しいと仮定すれば，データの周辺尤度 $p_i(\boldsymbol{x}_n)$ を最大にするモデル族を選択することになる．

ここで，データ $\boldsymbol{x}_n$ の周辺尤度は，ラプラス近似を用いて求めることができる．ラプラス近似とは，$\boldsymbol{\theta}$ が平均ベクトル $\widehat{\boldsymbol{\theta}}$，分散共分散行列 $J_q(\widehat{\boldsymbol{\theta}})^{-1}$ の p 次元の多変量正規分布の場合の，次式で表される近似式である．

$$\int \exp\{nq(\boldsymbol{\theta})\}d\boldsymbol{\theta} \approx \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}}|J_q(\widehat{\boldsymbol{\theta}})|^{\frac{1}{2}}} \exp\{nq(\widehat{\boldsymbol{\theta}})\} \tag{30}$$

この式の詳しい導出過程は省略するが，$q(\boldsymbol{\theta})$ を最尤推定量 $\widehat{\boldsymbol{\theta}}$ まわりでテイラー展開して導かれる．この近似式を用いると，周辺尤度 $p_i(\boldsymbol{x}_n)$ は以下のように近似される．

$$p_i(\boldsymbol{x}_n) = \int \exp\eta(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{31}$$

$$\approx \exp\eta(\widehat{\boldsymbol{\theta}})\pi(\widehat{\boldsymbol{\theta}}) \int \exp\left\{ -\frac{n}{2}\, {}^t(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})J_q(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right\}d\boldsymbol{\theta} \tag{32}$$

$$= \exp\eta(\widehat{\boldsymbol{\theta}})\pi(\widehat{\boldsymbol{\theta}})(2\pi)^{\frac{p}{2}}n^{-\frac{p}{2}}|J_q(\widehat{\boldsymbol{\theta}})|^{-\frac{1}{2}} \tag{33}$$

注意すべき点は，周辺尤度を対数尤度 $\eta(\boldsymbol{\theta})$ の定義を使って書き換えたのが，式 (31) であることと，式 (32) の近似式の導出には，$n \to +\infty$ のときに $\widehat{\boldsymbol{\theta}}$ が $\boldsymbol{\theta}$ に確率収束することと，収束のオーダーが $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{\frac{1}{2}})$ となることとを用いている．この近似の段階で，モデル族の選択の話から，パラメータをどのように選択するかということも考慮に入れたモデル選択の話に移行していることにも注意されたい．

式 (33) の対数をとり -2 を乗じると，

$$-2\log p(\boldsymbol{x}_n) \approx -2\eta(\widehat{\boldsymbol{\theta}}) + p\log n + \log|J_q(\widehat{\boldsymbol{\theta}})| - p\log(2\pi) - 2\log\pi(\widehat{\boldsymbol{\theta}}) \tag{34}$$

が得られる．ここで，データ数 n に関するオーダー $O(1)$ 以下の項を無視すると，

$$BIC \stackrel{\text{def}}{=} -2\eta(\widehat{\boldsymbol{\theta}}) + p\log n \tag{35}$$

が導出される．式 (35) の形からわかるように，BIC は AIC と同様，パラメータ数 p を考慮に入れたモデル選択基準になっているが，これは周辺尤度を近似する際に導出されたものである．

以上をまとめると，BIC の導出のための出発点となるベイズ統計学では，事後確率の大小がモデル選択の基準になるが，事後確率のうち周辺尤度だけがモデル選択にとって重要な因子である．この周辺尤度を近似して求めたのが，BIC である．

[*12] 母数空間 $\boldsymbol{\Theta}_i$ におけるすべての $\boldsymbol{\theta}_i$ に関するモデルの尤度を足し合わせていることからもわかるように，これがモデルの尤度ではなく，モデル族の尤度に相当することに注意されたい．

4.2 BIC における推論

AIC と同じように，BIC における推論を検討する．

(B1) 周辺尤度 $p_i(\boldsymbol{x}_n)$ を事前分布 $\pi_i(\boldsymbol{\theta})$ によってあらわす．

(B2) モデル族 M_i の事後確率 $P(M_i|\boldsymbol{x}_n)$ を周辺尤度 $p_i(\boldsymbol{x}_n)$ と事前確率 $P(M_i)$ とから求め，最大事後確率を与えるモデル族 M_i を選択する．

(B3) モデル族の事前確率 $P(M_i)$ を等しいと仮定すると，周辺尤度 $p_i(\boldsymbol{x}_n)$ の大小がモデル族のよさをあらわす．

(B4) 周辺尤度 $p_i(\boldsymbol{x}_n)$ は，漸近的に $BIC = -2\eta(\widehat{\boldsymbol{\theta}}) + p\log n$ に等しくなる．

まず注意しないといけない点は，(B1), (B2) における事前分布の設定である．赤池が BIC の問題点としてあげているのが，事前分布をどのように設定するかについての基準が存在しないことである．この点は，BIC 固有の問題というよりもむしろ，ベイズ統計学で用いられる推論に対し，よく提起される問題点である．

ベイズ統計学において，事前分布にはさまざまなものが想定されるが，どのような事前分布がいいのかに対する答えはない．事前分布を一意に定められないことが，ベイズ統計学に対する有効な反駁になっているかどうかに関しては，ここでは論じないことにして，赤池の解決法を簡単に述べたい．(B2) であらわれる $P(M_i)$ への事前分布の割り当てに関しては，BIC の導出と同様に，とりあえず (B3) を前提する．問題は，(B1) であらわれる $\pi_i(\boldsymbol{\theta}_i)$ への事前分布の割り当てで，これに関しては，事前分布が超母数と呼ばれるパラメータによってあらわされる確率分布であると仮定し，この超母数を最尤法によって推定する．これによって，事前分布はある基準に基づいて一意に定められる．こうして導出された情報量基準は，ABIC と呼ばれる．

BIC と赤池による ABIC とを比較したときに，BIC の場合，事前分布 $\pi(\boldsymbol{\theta})$ をどのようにすればいいのかについての規定はない．他方，ABIC の場合，事前分布 $\pi(\boldsymbol{\theta})$ は超母数 $\boldsymbol{\lambda}$ によって規定されている．そして，この $\boldsymbol{\lambda}$ を最尤法によって推定するので，$\pi(\boldsymbol{\theta}|\widehat{\boldsymbol{\lambda}})$ は尤度主義の立場に基づいて求められる．この意味で，赤池が懸念する，事前分布の設定をどうすればよいのかという疑念は払拭される．

しかし，BIC の問題点はそこにあるのではないように思われる．対数周辺尤度 $\log p_i(\boldsymbol{x}_n)$ 自体は，ベイズ統計学の考えに基づいた概念であるが，これから漸近的に求められた BIC はベイズ統計学の枠組みで説明できていないというのが，問題であるように思われる．すなわち，(B4) でなされる近似を行わなければ，BIC が導出されないのだが，この (B4) の過程が BIC の導出のすべてである．というのも，BIC の第 2 項で出てくる $p\log n$ は，(B4) の過程を経なければ，出てこないからである．しかし，この (B4) の過程をベイズ統計学の枠組みでどのように説明できるのだろうか．周辺尤度 $p_i(\boldsymbol{x}_n)$ 自体はベイズ統計学の枠組みで説明できる．しかし，標本数 n が「無限に大きい」場合にしか，$p_i(\boldsymbol{x}_n)$ の推定量 $-2\eta(\widehat{\boldsymbol{\theta}}) + p\log n$ は意味をなさないのだが，

この部分の，ベイズ統計学の枠組みでの説明が必要なのである．この意味で，AIC とまったく同じ問題を抱えていると言えよう．

5. モデル選択の問題

この章で，モデル選択基準として取り上げた AIC や BIC が採用する「よさ」について比較検討したいと思う．

まず，モデル選択とは何かについておさらいすると，準備された複数個のモデル族から，モデル選択基準に照らし合わせて，最適なモデルを選択することである．このようなモデル選択基準の候補として，本稿では，AIC と BIC の 2 つを取り上げた．ここで，情報量基準のうち，どれがモデル選択として適切な基準を提供するかについて，2 つ考えなければいけない点があるように思われる．1 つは，（ア）各情報量基準に照らし合わせて選択されたモデルが，実際に「よい」モデルかどうか，という実用上の問題である．もう 1 つは，（イ）情報量基準を導出する際に，そこで前提される概念が適切であるかどうかという，概念上の問題である．

まず（ア）についてであるが，これを確認するためには，モデルのもつ実用上の「よさ」が何かをはっきりさせる必要がある．AIC, BIC ともに，モデル選択の目的が，真のモデルを推定することではなく，（真のモデルに近いという意味で）よりよい予測を与えるモデルを選択することにあった，という点に注目されたい．この点に関しては，前章までの議論から理解されると思う．復習すると，AIC では真の分布 g を推定しているのでは決してなく，AIC が小さくなるものを準備されたモデルから選択しているのである．これは，BIC についても同じことが言える．この点をおさえれば，各情報量基準により得られたモデルが実際に現象の予測に成功しているならば，実用上「よい」モデルということになる．

そして，どの情報量基準が実際によいモデル選択基準を与えているかについては，まだ結論は出ていないように思われる．事実，AIC がモデルの選択に失敗している例がある．下平が取り上げる例は，複数個用意されたほ乳類の系統樹から，モデル選択基準に基づいて，もっともよい系統樹を選択する問題である[*13][*14]．

系統樹とは，共通祖先から始まってどのような順序で分岐したかをあらわす「樹」のことである．下平は，6 種類のほ乳類（ヒト，アザラシ，ウシ，ウサギ，マウス，オポッサム）に限定して，それらの系統樹の作成を行っている．データとして用いられるのは，これらのほ乳類から抽出したミトコンドリア DNA のうち，タンパク質を表現している遺伝子の塩基配列だけを，アミノ酸配列に変換したものである．そして，オポッサムの祖先から，他のほ乳類が分岐したという（生物学で認められている）事前情報をもとに，すべての可能な系統樹を作成する．このとき，各系統樹に対し，アミノ酸配列データの確率モデルを特定することができるので，モデル選択基

[*13] 詳しくは，[下平 2004], [赤池 2007] を参照されたい．

[*14] ここでの系統樹についての記述に関して，田中泉吏氏（京都大学）から有益なコメントをいただき，この場を借りて感謝を申し上げます．

準によって，先ほどのデータを用いて，モデル選択が行える．

しかし，AIC に基づいてモデル選択を行った場合に，従来はネズミの祖先からウサギが分岐したと考えられていたのに対し，ウサギの祖先からヒトが分岐したという（従来の生物学の考えを覆す）「発見」がなされた．しかし，さらに大量のデータを用いると，従来の生物学の考えと合致する帰結が得られる．このように，AIC が万能なモデル選択基準であるとは言えない[*15]のだが，AIC がモデル選択基準として成功している事例も多数あり，実用面ではどの情報量基準がよりよいモデル選択基準を与えるのかについての解答は出ていないように思われる．

次に，（イ）についてであるが，これを確認するためには，各情報量基準を導出する際に用いられている推論において前提とされている概念の適切さ，たとえば整合的であるかや節約の原理にのっとったものであるかどうかなど，を比較検討しなければならない．

AIC の導出の際に，カルバック–ライブラー情報量と切り離して考えなければならない点について論じた．すなわち，カルバック–ライブラー情報量に関してだけ見れば，平均尤度に基づいて，モデルの良さが判定される．しかし，AIC に関して言えば，最尤法だけに基づいて導出されたのではなく，漸近的に導出されたバイアス $b(g)$ が含まれている．この $b(g)$ については，尤度主義の枠組みでは，概念的基礎は与えられない．前章で述べたことの繰り返しになるが，頻度論的な意味での「よさ」の概念が「こっそり」もちこまれているのであり，これら別々の概念的基礎のもとで，AIC が成り立っており，この点を説明することを尤度主義の枠組みだけから行うことができないのである．BIC についても同様で，ベイズ統計学の枠組みだけで，「よさ」の基準が求められているのではない．このように，2 つの情報量基準はどちらも，異なる概念的基礎の上に成立しており，AIC のように尤度原理を出発点として導出されたモデル選択を行うべきか，あるいは BIC のように事後確率の比較を出発点として導出されたモデル選択を行うべきかという問題の解決のためには，これらの情報量基準の導出の前提となる統計的推論（尤度主義に基づいたものか，あるいは，ベイズ主義に基づいたものか）のどちらが概念的により適切であるかという問題の解決以前に，答えるべき困難な問題が残されているように思われるのである[*16]．

6. まとめ

モデル選択基準として，AIC と BIC という 2 つの情報量基準がどのような推論に基づいて導出されるかを考察してきた．2 つの情報量基準ともに，統計的モデルの「よさ」は，真のモデルとの適合度とパラメータ数の両方で判定されることをみてきた．しかし，これらの情報量基準

[*15] この点については，慎重に議論をしなければならない．というのも，統計的推論である以上，誤った推論になる可能性は，その確率が小さいにせよ，ありうるのである．この点をどう実用上の「よさ」の議論と結びつけるかであるが，1 つは [下平 2004] のように，その確率が（統計学者の判断で）大きいと考えられる場合には，実用上問題のある統計的手法であると言える，と思われる．

[*16] このようなモデル選択基準についての，概念的な分析に基づいたよさに関する議論は，[Forster 2004] において Forster らと Boik とのあいだで行われている．しかし両者ともに，各情報量基準の導出において異なる概念的基礎の混在（つまり，頻度論的な考え方が混入していること）がなかったかのような議論を行っており，このような議論は不毛であると私は考える．

が導出される過程で，その概念的基礎に問題があることも同時にわかった．その一方で，これらの情報量基準が実用的に統計的モデルの決定に成功してきたという事実もある．この意味では，実用的には「よい」モデルの発見に，情報量基準が寄与しているとも言える．本論文では，統計的モデルの「よさ」を情報量基準の導出に焦点を当てて検討を行った関係で，「よさ」についての概念的側面に偏った議論になってしまった．しかし，実用的な「よさ」もまた科学において重要な側面であり，この点を含めた統計的モデルの「よさ」についての検討が今後の課題だと言えよう．

参考文献

[Akaike 1973] Akaike, H. (1973) , “Information Theory and an Extension of the Maximum Likelihood Principle”, *2nd international symposium on information theory*, 243-247.

[Akaike 1974] Akaike, H. (1974) , “A New Look at the Statistical Model Identification”, *IEEE Transactions on Automatic Control*, **19**, 716-723.

[Akaike 1978a] Akaike, H. (1978) , “A Bayesian Analysis of the Minimum AIC Procedure”, *Ann. Inst. Statist. Math.*, **30**, Part A, 9-14.

[Akaike 1978b] Akaike, H. (1978) , “A New Look at the Bayes Procedure”, *Biometrika*, **65**, 1, 53-59.

[Akaike 1980] Akaike, H. (1980) , “Likelihood and the Bayes Procedure”, In *Bayesian Statistics*, eds. Lindley, D. V. *et al*, 1-13.

[Akaike 1982] Akaike, H. (1982) , “On the Fallacy of the Likelihood Principle”, *Statistics & Probability Letters*, **1**, 75-78.

[Edwards 1992] Edwards, A. W. F. (1992) , *Likelihood*, Expanded Edition, the Johns Hopkins University Press.

[Schwarz 2000] Schwarz, G. (2000) , “Estimating the Dimension of a Model”, *The Annals of Statistics*, Vol. **6**, No.2, 461-464.

[Forster 1994] Forster, M. & Sober, E. (1994) , “How to Tell when Simpler, More Unified, or Less Ad Hoc Theory will Provide More Accurate Predictions”, *British Journal for the Philosophy of Science*, **45**, 1-35.

[Forster 2000] Forster, M. (2000) , “Key Concepts in Model Selection: Performance and Generalizability”, *Journal of Mathematical Psycology*, **44**, 205-231.

[Forster 2004] Forster, M. & Sober, E (2004) , “Why Likelihood?”, In *The Nature of Statistical Evidence*, eds. Taper, M. L. & Lele S. P, 153-190.

[Hacking 1965] Hacking, I. (1965) , *Logic of Statistical Inference*, Cambridge University Press.

[Zucchini 2000] Zucchini, W. (2000) , “An Introduction to Model Selection”, *Journal of*

Mathematical Psycology, **44**, 41-61.

[赤池 2007] 赤池弘次他 (2007)，赤池情報量基準 AIC，共立出版.

[内井 1995] 内井惣七 (1995)，科学哲学入門，世界思想社.

[坂元 1983] 坂元慶行他 (1983)，情報量統計学，共立出版.

[小西 2004] 小西貞則・北川源四郎 (2004)，情報量基準，朝倉書店.

[下平 2004] 下平英寿他 (2004)，モデル選択 予測・検定・推定の交差点，岩波書店.

[竹村 1991] 竹村彰通 (1991)，現代数理統計学，創文社.