

Title	フィッシャーの情報量
Sub Title	Fisher's amount of information
Author	竹中, 淑子(Takenaka, Yoshiko) 金川, 秀也(Kanagawa, Shuya)
Publisher	慶應義塾経済学会
Publication year	2005
Jtitle	三田学会雑誌 (Keio journal of economics). Vol.98, No.1 (2005. 4) ,p.95- 104
JaLC DOI	10.14991/001.20050401-0095
Abstract	
Notes	研究ノート
Genre	Journal Article
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN00234610-20050401-0095

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

研究ノート

フィッシャーの情報量

竹 中 淑 子
金 川 秀 也

（初稿受付 2004 年 11 月 18 日，
査読を経て掲載決定 2005 年 4 月 19 日）

はじめに

この研究ノートの目的は、「フィッシャーの情報量」の起源から現代の情報量に至るまでのその概念の果たした役割の重要性を明らかにすることである。そこで「フィッシャーの情報量」，「エントロピー」，「情報量規準 AIC 」の関係に言及し，特に最尤推定量から AIC がどのように導かれたかを詳細，厳密に解説してみた。ここでも，フィッシャーの情報量の重要な性質が， AIC の理論的根拠になっていることを見ることができる。

1. なぜフィッシャーの情報量なのか

著者の一人は数年前から統計的推定検定に関する R. A. フィッシャー（1890–1962）と J. ネイマン（1894–1981）との「フィッシャー＝ネイマン論争」といわれる論争に興味を持ち，それに関するいくつかの著作，論文などを読んで

できた。論争の主な論点の 1 つは点推定における対立仮説，検定力関数に関するものと，もう 1 つは区間推定における推測確率あるいはフィッシャーの験信度（fiducial probability）に関するものであった。多くの統計学者をまきこんでのフィッシャーの死後も続いた 30 年に及ぶ論争であったが，1970 年代後半には決着した感がある。1978 年の国際統計学会総会で，J. G. ペーダセンが包括的に整理された総合報告を行ったが，それには，

——R. A. フィッシャーの数多くの重要な統計学への貢献の中にあつて，フィッシャー験信論は極めて局限された成功しか実現できなかったし，現在では本質的に死滅している——

となっていた [5]。このように，この論争に関する限り，フィッシャーに分はなかったようだし，また，それとは別に統計学の大部分はネイマン流に定式化され記述されてきた感がある。それは統計学の論文や専門書を読んでいると，特にことわって，フィッシャーの意

味での何々（たとえば、フィッシャーの意味での最良検定法）やフィッシャーの何々の定義（たとえば、フィッシャーの有効性の定義、一致性の定義）などの記述に出会うことからわかる。また、「フィッシャー＝ネイマン論争」についての北川敏男（1909–1993）の論述に、1988年の時点度として、次のような文章がある [6]。

——統計学最前線の研究を見ると、情報量理論、多変量解析論、情報幾何学、条件付尤度論等の何れにおいても、起源はフィッシャーの着想にある。これに反し、1950年代から20年余りにわたって主流をなしたネイマン－ピアソン流のゆき方は今では湧き出る発想の泉の役を果たしていない——

そこで、この文章の一端を示すべく、フィッシャーにより導入された数多くの概念、理論、方法の中から「フィッシャーの情報量」をえらび、起源から、現代の情報量による推定量に至るまでのその概念の果たした重要性を探ってみた。

2. フィッシャーの情報量

未知のパラメータ θ を含む確率密度関数 $f(x; \theta)$ をもつ母集団から大きさ n の独立無作為標本 (x_1, x_2, \dots, x_n) をとるとき、 $s(x_i) = \frac{\partial}{\partial \theta} (\log f(x_i; \theta))$, $i = 1, 2, \dots, n$ をフィッシャーはスコア (score) といった (1921年)。これは θ の関数とすると、 x_i が与えられたとき θ が少し変化したときの関数 f の相対的な変化を表す。 $s(x_i)$ に対応する確率変数 $s(X_i)$ についてみると、スコア (X_i) の期待値 $E\{s(X_i)\}$ は 0 で、分散 $V\{s(X_i)\}$ は

$$\begin{aligned} V\{s(X_i)\} &= E\left[\left(\frac{\partial}{\partial \theta} \log f(x_i; \theta)\right)^2\right] \\ &= -E\left\{\frac{\partial^2}{\partial \theta^2} \log f(X_i; \theta)\right\} \end{aligned}$$

となる。このスコアの分散を内在精度 (intrinsic accuracy) といった。大きさ n の標本 (x_1, x_2, \dots, x_n) をとったとき、 θ に関する尤度関数

$$L_n(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

の対数をとった $L_n(\theta)$ については

$$\begin{aligned} E\left[\left(\frac{\partial}{\partial \theta} \log L_n(\theta)\right)^2\right] \\ = E\left\{-\frac{\partial^2}{\partial \theta^2} \log L_n(\theta)\right\} \end{aligned}$$

となる。これを $I(\theta)$ と表し、大きさ n の標本 (x_1, x_2, \dots, x_n) にもとづく θ に関するフィッシャーの情報量 (amount of information) とのちに言われることになる。ここでは F -情報量と略記する。

パラメータが k 個ある場合には、 F -情報量は $k \times k$ 行列となる。すなわち、大きさ n の標本 (x_1, x_2, \dots, x_n) にもとづく未知パラメータ $(\theta_1, \theta_2, \dots, \theta_k) = \theta$ の F -情報行列 $I(\theta)$ とは、第 (i, j) 成分を

$$\begin{aligned} I_{ij} &= E\left[\left(\frac{\partial}{\partial \theta_i} \log L_n(\theta)\right)\left(\frac{\partial}{\partial \theta_j} \log L_n(\theta)\right)\right] \\ &= E\left(-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L_n(\theta)\right), \quad 1 \leq i, j \leq k \end{aligned}$$

とする行列である。任意の統計量 $\mathbf{T} = T(x_1, \dots, x_n)$ に含まれる θ についての F -情報量を $I_{\mathbf{T}}(\theta)$ とすると、

$$I_{\mathbf{T}}(\theta) \leq I(\theta)$$

となる。そこでフィッシャーは $\frac{I_T(\theta)}{I(\theta)}$ (≤ 1) を有限の n についての推定量の効率と定義した (1925 年)。さらに

$$\lim_{n \rightarrow \infty} \frac{I_T(\theta)}{I(\theta)}$$

を大標本における推定量の漸近効率と定義した。この概念はのちに漸近有効性の議論に発展する。 $I_T(\theta) = I(\theta)$ のとき、推定量と標本の情報は同じで、情報の損失はない。こういう推定量が十分推定量である。フィッシャーは、情報量ゼロ、情報の回復 (喪失した情報の一部回復)、情報程度 (informativeness) など論じている。 F -情報量に類するものは他にいくつかある。M.G. Kendall は

$$E \left\{ \left(\frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \right\} \\ = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \theta} \log f(x; \theta) \right)^2 f(x; \theta) dx$$

を θ に関する情報がどれだけ得られるかという意味で、標本 1 つあたりの報知高といった。また、伊藤清は母集団分布が m 個のパラメータ $\theta_1, \theta_2, \dots, \theta_m$ を含むとき、 $\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_k$ を統計量とし、その分布密度を $g(t_1, t_2, \dots, t_k; \theta_1, \theta_2, \dots, \theta_m)$ として、フィッシャーの定義を拡張して

$$I(\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_k) \\ = E \left\{ \sum_{i=1}^m \frac{\partial}{\partial \theta_i} \log g(\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_k; \theta_1, \theta_2, \dots, \theta_m) \right\}$$

を $\theta_1, \theta_2, \dots, \theta_m$ に関する統計量 $\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_k$ の報知高 (amount of information) といった。パラメータの変化が統計量の分布に平均的にどれだけ影響を及ぼすかを表す量である。

3. パラメータの微小変化と F -情報量

F -情報量の定義は対数尤度の 2 階微分の平均値であるから、特に情報量を意識しない場合にも無意識に使われていることもある。ここでは、情報の尺度としての本来の意味で、情報の損失についてみたときやパラメータの微小変化に対する F -情報量の性質を挙げてみる。

3.1 情報の限界を表す式

確率密度関数 $f(x; \theta)$ をもつ母集団からの標本 (x_1, x_2, \dots, x_n) に対する θ の推定量 $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ が、 θ の不偏推定量すなわち $E(\hat{\theta}) = \theta$ とすると、その分散 $V(\hat{\theta})$ について

$$V(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

が成立する。クラメル=ラオの不等式として知られる式である。情報の限界を表す式とみなすことができ、この不等式で等号に等しい分散をもつ不偏推定量が最小分散不偏推定量 (有効推定量) である。

パラメータが $m (\geq 2)$ 個ある場合、クラメル=ラオの不等式は次のようになる。 $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ の不偏推定量を $\hat{\theta}$ とし、その $\hat{\theta}$ の分散共分散行列を $V(\hat{\theta})$ 、 F -情報行列 $I(\theta)$ の逆行列を $I(\theta)^{-1}$ とすると、適当な正則条件の下で、

$$V(\hat{\theta}) \geq I(\theta)^{-1}$$

となる。つまり $V(\hat{\theta}) - I(\theta)^{-1}$ は半正値定

符号行列となる。

3.2 分布の内在精度の尺度としての情報

F -情報量 $I(\theta)$ は、パラメータの値の微小変化に関する確率変数の感度の尺度であり、分布の内在精度の尺度としての情報量（未知のパラメータについての不確実性が減少する度合い）を表すとみなされる。さまざまに定義される分布間の距離において、 $I(\theta)$ が、パラメータ θ の値の微小変化に関する確率変数の感度の尺度となることがわかる。ヘリンガーの距離もその1つである。

ヘリンガーの距離とは、パラメータの θ' についての X の確率密度を $p(x; \theta)$, $p(x; \theta')$ とするとき、2つの分布間を距離関数

$$H_p(\theta, \theta') = \cos^{-1} \int_{-\infty}^{\infty} \sqrt{p(x; \theta)p(x; \theta')} dx$$

で測るものである。 $\theta = \theta'$ ならばヘリンガーの距離は

$$\begin{aligned} H_p(\theta, \theta') &= \cos^{-1} \int_{-\infty}^{\infty} p(x; \theta) dx \\ &= \cos^{-1} 1 = 0 \end{aligned}$$

また、 $p(x; \theta)$ と $p(x; \theta')$ が直交する場合最大となり、

$$H_p(\theta, \theta') = \cos^{-1} 0 = \frac{\pi}{2}$$

である。 $\theta' = \theta + \Delta\theta$ として、 $p(x; \theta')$ をテーラー展開し、 $\Delta\theta$ の高次の項を無視すると、

$$\begin{aligned} H_p(\theta, \theta') &\cong \cos^{-1} \int_{-\infty}^{\infty} p(x; \theta) \left\{ 1 - \frac{1}{8} \left(\frac{p'(\theta)}{p(\theta)} \right)^2 (\Delta\theta)^2 \right\} dx \\ &= \cos^{-1} \left\{ 1 - \frac{1}{8} I(\theta) (\Delta\theta)^2 \right\} \end{aligned}$$

となる。 $I(\theta)$ は正であるから、 $I(\theta)$ の値が増加すればヘリンガーの距離も増加する。その

意味で $I(\theta)$ はパラメータの値の微小変化に関する確率変数の感度の尺度となっている。

パラメータが $m \geq 2$ 個の場合、2つの分布間のヘリンガー距離は $I(\theta) (d\theta)^2$ の代わりに2次微分形式

$$\sum_{i=1}^m \sum_{j=1}^m I_{ij} \Delta\theta_i \Delta\theta_j$$

となり、パラメータの変化に関する確率変数の感度は、 F -情報量行列 $I(\theta)$ に依存することがわかる。

4. 最尤法

4.1 最尤法の理論的根拠

確率変数が $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ をパラメータとして含む確率密度関数 $f(x; \theta)$ をもち、 X の標本 (x_1, x_2, \dots, x_n) をとったとき、 θ の尤度

$$L(\theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta)$$

を最大にする θ の値 $\hat{\theta}$ をパラメータ θ の推定量とするのが最尤法で、 $\hat{\theta}$ を最尤推定量といった。なぜ最大にするのがよいかは、「現実起きた事象はその可能性（確率）が大きいはずである」という統計の基本理念に従ったものだが、理論的根拠がはっきりされないまま使われた感もある。

最尤推定量には、次のような漸近正規性という大変良い性質がある。標本 (x_1, x_2, \dots, x_n) が連続な確率密度関数 $f(x; \theta)$ によって分布するとき最尤推定量 $\hat{\theta}$ は $n \rightarrow \infty$ のとき、近似的に、平均 0、分散 $I(\theta)^{-1}$ の正規分布に従う。

$$\hat{\theta} \sim N(0, I(\theta)^{-1})$$

パラメータ $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ のときも最尤推定量 $\hat{\theta}$ は $n \rightarrow \infty$ のとき、近似的に、平均 0, 分散共分散行列 $I(\theta)^{-1}$ の正規分布に従う。ここに $I(\theta)$ はフィッシャーの情報行列である。

この漸近正規性より、最尤推定量が一致性 ($\hat{\theta}$ は θ に収束する), 漸近不偏性, 漸近有効性 (分散最小の不偏推定量) などの良い性質をもつことがわかる。さらに最尤推定量の漸近正規性より、最尤推定量にもとづく検定は漸近的には最良の検定となることが導かれる。帰無仮説 $\theta = \theta_0$ に対し

$$\hat{\theta} \geq \theta_0 + \frac{1}{\sqrt{I(\theta)}} K_{2\alpha}$$

を棄却域にする片側検定である。

しかし、最尤法の理論的根拠となると、それは分布間の距離、情報量、エントロピーなどの概念を使うとはっきりすることを以下に述べる。最尤推定量 $\hat{\theta}$ は $L(\theta)$ ではなく、その対数をとった対数尤度 $l(\theta)$

$$l(\theta) = \log L(\theta) = \sum_{i=1}^n \log f(x_i; \theta)$$

を最大とする θ の値ということになる。ここで、エントロピー、 KL -情報量との関連がでくる。2つの確率密度関数 $f(x; \theta^*)$ と $f(x; \theta)$ に対するエントロピー $B(\theta^*, \theta)$ は、

$$B(\theta^*, \theta) = \int_{-\infty}^{\infty} f(x; \theta^*) \log f(x; \theta) dx - \int_{-\infty}^{\infty} f(x; \theta^*) \log f(x; \theta^*) dx$$

であり、0 または負の値をとる。 θ^* を真のパラメータ ($f(x; \theta^*)$ が真の確率密度関数) とするとき、 $B(\theta^*, \theta)$ が大きいほど、すなわち

$$\int_{-\infty}^{\infty} f(x; \theta^*) \log f(x; \theta) dx \quad \dots (*)$$

が大きいほど、 θ は θ^* の良い近似を与える。上記の最尤推定量 $\hat{\theta}$ は大数の法則により、(*) の最大値の一致推定量となり、これが最尤推定量の理論的根拠である。また、エントロピー $B(g: f(x; \theta))$ の符号を反転した値が KL -情報量、つまり

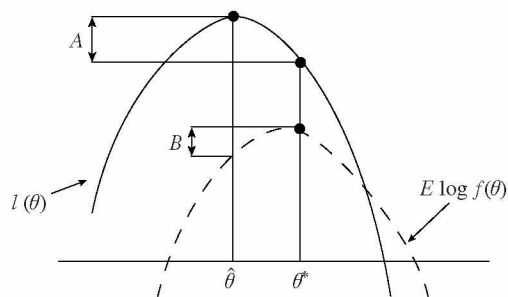
$$KLI(g: f(x; \theta)) = -B(g: f(x; \theta))$$

であるから、最尤推定量 $\hat{\theta}$ は KLI を最小にする θ であるともいえる。これで最尤推定量の意味が明確になる。

4.2 最尤法から AIC へ

最尤推定量よりさらに良い推定量はないか、あるいは最大尤度が同程度の場合はどうするかと考えられてきた。平均対数尤度 $\frac{1}{n} l(\theta)$ が、 $E[\log f(x; \theta)]$ の推定値として使えることはすでにみた。標本から推定された評価基準である $l(\theta)$ の最大値 $\hat{\theta}$ と、本当の評価基準であるべき $E[\log f(x; \theta)]$ の最大値 θ^* との差をみる。

図のように、 $l(\theta)$ で評価すると、 $\hat{\theta}$ は θ^* より A だけ良くみえるが、 $E[\log f]$ で評価すると、 $\hat{\theta}$ は θ^* より B だけ悪い。したがって、



図

$$l(\theta) - (A + B)$$

が、 $E[\log f(x; \hat{\theta})]$ の推定値すなわち $\hat{\theta}$ の本当の良さになる。実は、 $A + B$ はパラメータ数の近似を与える。したがって、

$$l(\theta) - k$$

が $E[\log f(x; \hat{\theta})]$ の推定値、すなわち $\hat{\theta}$ の本当の良さになる。これを -2 倍したものが、情報量規準 AIC (Akaike Information Criterion) である。すなわち、

$$AIC(k) = -2\ell(\hat{\theta}) + 2k$$

AIC は小さければよいわけで、 AIC を最小にする推定量を最小 AIC 推定量という。最大尤度が同程度の場合は、パラメータ数が最も少ないものを選べばよいということになる。そういう意味で最小 AIC 推定量は最尤推定量を一步進めたものであると考えられる。ここでは、大雑把に AIC の導き方を述べたが、次章では詳細に KL -情報量から AIC の導き方を解説する。 F -情報量、最尤法の前述の性質が最小 AIC 推定量の理論的根拠になっていることを見ることができる。

5. AIC とフィッシャーの情報量

坂本, 石黒, 北川 [7] に従い、 AIC とフィッシャーの情報量の関係について解説する。

5.1 Fisher (フィッシャー) の情報量と Kullback-Leibler (カルバック・ライブラー) 情報量

$f(x, \theta)$ をパラメータ θ とする密度関数とする、ただし θ は 1 次元の未知母数。無作

為抽出による標本 $\mathbf{x} = (x_1, x_2, \dots, x_n)$ に対応する密度関数を $f(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta)$ とおく。このとき尤度関数 $L(\theta) = L(\mathbf{x}, \theta) = \log f(\mathbf{x}, \theta)$ 及び、 $f(x, \theta)$ を密度関数とする n 次確率変数 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ に対して

$$\begin{aligned} I(\theta) &= -E \left\{ \frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right\} = -E \left\{ \frac{\partial^2 \log L(\mathbf{X}; \theta)}{\partial \theta^2} \right\} \\ &= - \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\partial^2 \log f(\mathbf{x}; \theta)}{\partial \theta^2} \prod_{i=1}^n f(x_i; \theta) dx_1 \dots dx_n \end{aligned}$$

がフィッシャーの情報量である。さらに $\theta = (\theta_1, \theta_2, \dots, \theta_K)$ が多次元の場合は

$$\begin{aligned} I_{ij}(\theta) &= -E \left\{ \frac{\partial^2 \log L(\theta)}{\partial \theta_i \partial \theta_j} \right\} \\ &= - \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\partial^2 \log f(\mathbf{x}; \theta)}{\partial \theta_i \partial \theta_j} \prod_{i=1}^n f(x_i; \theta) dx_1 \dots dx_n \end{aligned}$$

を各要素とする $K \times K$ 行列 $I(\theta)$ がフィッシャーの情報行列である。後で定義される $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ の最尤推定量 $\hat{\theta}_k$ に対して漸近正規性 (中心極限定理)

$$\sqrt{n}(\hat{\theta}_k - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}) \quad (n \rightarrow \infty)$$

が成り立ち、その極限分布である正規分布の共分散行列がフィッシャーの情報行列の逆行列となる。このことが AIC の理論的根拠となる。

2 つの離散分布 $\mathbf{p} = (p_1, p_2, \dots, p_m)$, $\mathbf{q} = (q_1, q_2, \dots, q_m)$ について \mathbf{p} を真の分布、 \mathbf{q} をその 1 つのモデルとする。このとき

$$\begin{aligned} KLI(\mathbf{p}; \mathbf{q}) &= \sum_{i=1}^m p_i \log \frac{p_i}{q_i} \\ &= \sum_{i=1}^m p_i \log p_i - \sum_{i=1}^m p_i \log q_i \end{aligned}$$

がモデル \mathbf{q} に対する真の分布 \mathbf{p} のカルバック・ライブラー情報量 (KL -情報量) で、 \mathbf{p}, \mathbf{q}

間の距離を表す。真の分布 \mathbf{p} は固定されているから $\sum_{i=1}^m p_i \log q_i$ が大きいほど KL -情報量 $KLI(\mathbf{p}; \mathbf{q})$ が小さくなり、2つのモデル \mathbf{p}, \mathbf{q} は近いといえる。

5.2 離散型分布モデル \mathbf{q} の対数尤度と平均対数尤度

$\mathbf{q} = (q_1, q_2, \dots, q_m)$ を値 (z_1, z_2, \dots, z_m) に対応する離散型分布の1つのモデルとする。 \mathbf{q} が真の分布に適合しているか検定するために n 個の標本を母集団から取り出したとする。この n 個の標本の中で (z_1, z_2, \dots, z_m) に対応する標本の大きさを $\{n_1, n_2, \dots, n_m\}$ とする。すなわち、 $n = n_1 + n_2 + \dots + n_m$ 。このとき

$$\ell(\mathbf{q}) = \sum_{i=1}^m n_i \log q_i$$

はモデル \mathbf{q} の対数尤度である。 $\mathbf{p} = (p_1, p_2, \dots, p_m)$ を真の分布とすると大数の法則から確率1で

$$\lim_{n \rightarrow \infty} \frac{n_i}{n} = p_i, \quad 1 \leq i \leq m$$

が成り立つ。ゆえに

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ell(\mathbf{q}) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^m n_i \log q_i = \sum_{i=1}^m p_i \log q_i, \quad w.p.1$$

ここで $\sum_{i=1}^m p_i \log q_i$ は \mathbf{q} の平均対数尤度である。また $w.p.1$ は「確率1」を表す。

5.3 最尤モデルの平均対数尤度

ここからは母集団は連続分布に従い、その確率密度関数を $f(x; \theta)$ とする。ただし、

$$\theta = (\theta_1, \theta_2, \dots, \theta_K) \in \Theta_K$$

ここで Θ_K は自由パラメータ数（自由度） K のパラメータ空間とする。また

$$\theta^* = (\theta_1^*, \theta_2^*, \dots, \theta_K^*) \in \Theta_K$$

を真のパラメータとする。 n 個の標本 (x_1, x_2, \dots, x_n) に対して

$$\ell(\theta) = \sum_{i=1}^n \log f(x_i; \theta)$$

が対数尤度である。ここで $\log f(x_i; \theta)$ は K 個のパラメータ $\theta = (\theta_1, \theta_2, \dots, \theta_K) \in \Theta_K$ で規定される密度関数で、このような分布（確率密度関数）全体を $MODEL(K)$ とする。すなわち

$$MODEL(K)$$

$$= \{f(x; \theta) \mid \theta = (\theta_1, \theta_2, \dots, \theta_K) \in \Theta_K\}$$

$\ell(\theta)$ の最大値を最大対数尤度という。 $\ell(\theta)$ が最大ということは、 θ^* と θ をそれぞれパラメータとする2つの密度関数 $f(x; \theta^*)$, $f(x; \theta)$ の KL -情報量 $KLI(\theta^*; \theta)$

$$KLI(\theta^*; \theta)$$

$$= \sum_{i=1}^n f(x_i; \theta^*) \log \frac{f(x_i; \theta^*)}{f(x_i; \theta)}$$

$$= \sum_{i=1}^n f(x_i; \theta^*) \log f(x_i; \theta^*) - \sum_{i=1}^n f(x_i; \theta^*) \log f(x_i; \theta)$$

が最小になる。そこで最大対数尤度

$$\ell(\hat{\theta}_K) = \max_{\theta \in \Theta_K} \ell(\theta)$$

によって定義される推定量 $\hat{\theta}_K$ を θ^* の最尤推定量という。標本 (x_1, x_2, \dots, x_n) に対する分布（確率密度関数） $f(\cdot; \theta)$ の平均対数尤度を $\frac{1}{n} E \ell(\theta)$ とする。このとき確率変数 X の従う分布は $f(x; \theta^*)$ を確率密度関数とすることから

$$\frac{1}{n} E \ell(\theta)$$

$$= \frac{1}{n} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \sum_{i=1}^n \log f(x_i; \theta) \prod_{j=1}^n f(x_j; \theta^*) dx_1 \dots dx_n$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log f(x_i; \theta) \prod_{j=1}^n f(x_j; \theta^*) dx_1 \cdots dx_n \\
&= \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} \log f(x_i; \theta) f(x_i; \theta^*) dx_i \\
&= \int_{-\infty}^{\infty} \log f(x; \theta) f(x; \theta^*) dx = E[\log f(X; \theta)].
\end{aligned}$$

$E[\log f(X; \theta)]$ が平均対数尤度の本来の定義である。大数の法則から

$$\begin{aligned}
&\lim_{n \rightarrow \infty} \frac{1}{n} \ell(\theta) \\
&= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log f(x_i; \theta) = E[\log f(X; \theta)], \quad w.p.1
\end{aligned}$$

が成り立つ。この式は対数尤度が平均対数尤度の一致推定量であることを示している。平均対数尤度の n 倍を $\ell^*(\theta)$ とする。

$$\begin{aligned}
\ell^*(\theta) &= n \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{i=1}^n \log f(x_i; \theta) \prod_{j=1}^n f(x_j; \theta^*) dx_1 \cdots dx_n \\
&= n E[\log f(X; \theta)]
\end{aligned}$$

$\ell^*(\theta)$ が大きいほど θ の真の分布 θ^* に対する近似が良いことになる。最尤モデル $f(\cdot; \hat{\theta}_K)$ の平均対数尤度、すなわち最大対数尤度の n 倍が $\ell^*(\hat{\theta}_K) = n\ell(\hat{\theta}_K)$ である。 $\ell^*(\hat{\theta}_K)$ の期待値を最尤モデル $f(\cdot, \hat{\theta}_K)$ の期待平均対数尤度 $\ell_n^*(K)$ という。すなわち

$$\begin{aligned}
\ell_n^*(K) &= E[\ell^*(\hat{\theta}_K)] \\
&= E\left[n \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{i=1}^n \log f(x_i; \hat{\theta}_K) \prod_{i=1}^n f(x_i; \theta^*) dx_1 \cdots dx_n\right]
\end{aligned}$$

期待平均対数尤度 $\ell_n^*(K)$ は個々の実現値（標本値）に依存しない量であり、その値が大きいほどモデルの近似が良いといえる。

5.4 MODEL(K) に対する AIC

上述のように期待平均対数尤度 $\ell_n^*(K)$ は個々の実現値（標本値）に依存しない量であ

り、その値が大きいほどモデルの近似が良いといえる。しかし現実には母集団の真の分布がわからないため $\ell_n^*(K)$ を求めることが容易ではない。本セクションでは期待平均対数尤度 $\ell_n^*(K)$ の近似を与え、この近似式を用いて AIC を導く。

関数 $\ell^*(\theta)$ にパラメータ $\theta = \theta^*$ の周りでテーラーの定理を適用し、 $\theta = \theta^*$ を代入して次の近似式を得る。

$$\begin{aligned}
\ell^*(\hat{\theta}_K) &= \ell^*(\theta^*) + (\hat{\theta}_K - \theta^*) \left. \frac{\partial \ell^*(\theta)}{\partial \theta} \right|_{\theta=\theta^*} \\
&\quad + \frac{1}{2} (\hat{\theta}_K - \theta^*) \left. \frac{\partial^2 \ell^*(\theta)}{\partial \theta^2} \right|_{\theta=\theta^*} (\hat{\theta}_K - \theta^*)^T + o\left(|\hat{\theta}_K - \theta^*|^3\right)
\end{aligned}$$

式を簡単にするために $\log f(\mathbf{x}; \hat{\theta}_K) = \sum_{i=1}^n \log f(x_i; \hat{\theta}_K)$ とおく。 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ を用いて $\ell^*(\theta) = nE[\log f(\mathbf{X}; \theta)]$ と書くことができるので上記の近似式は

$$\begin{aligned}
\ell^*(\hat{\theta}_K) &= \ell^*(\theta^*) + n(\hat{\theta}_K - \theta^*) E\left[\left. \frac{\partial \log f(\mathbf{X}; \theta)}{\partial \theta} \right|_{\theta=\theta^*}\right] \\
&\quad + \frac{1}{2} n(\hat{\theta}_K - \theta^*) E\left[\left. \frac{\partial^2 \log f(\mathbf{X}; \theta)}{\partial \theta^2} \right|_{\theta=\theta^*}\right] (\hat{\theta}_K - \theta^*)^T + o(1)
\end{aligned}$$

まず $\theta = \theta^*$ で $\ell^*(\theta) = E[\log f(\mathbf{X}; \theta)]$ が最大値をとることから右辺第 2 項は 0 となる。

次に右辺第 3 項の期待値の部分を

$$I(\theta^*) = -E\left[\left. \frac{\partial \log f(\mathbf{X}; \theta)}{\partial \theta} \right|_{\theta=\theta^*}\right]$$

とおく。このとき $I(\theta^*)$ はフィッシャー情報行列で、各成分は

$$\begin{aligned}
I_{ij}(\theta^*) &= -E\left[\left. \frac{\partial \log f(\mathbf{X}; \theta)}{\partial \theta_i} \frac{\partial \log f(\mathbf{X}; \theta)}{\partial \theta_j} \right|_{\theta=\theta^*}\right], 1 \leq i, j \leq K
\end{aligned}$$

ゆえに

$$\begin{aligned} & \frac{1}{2}n(\hat{\theta}_K - \theta^*)^T E \left[\frac{\partial^2 \log f(X; \theta)}{\partial \theta^2} \right] \Big|_{\theta=\theta^*} (\hat{\theta}_K - \theta^*) \\ &= -\frac{1}{2}\sqrt{n}(\hat{\theta}_K - \theta^*)^T I(\theta^*) \sqrt{n}(\hat{\theta}_K - \theta^*) \end{aligned}$$

最尤推定量 $\hat{\theta}_K$ の漸近正規性から

$$\sqrt{n}(\hat{\theta}_K - \theta^*) \xrightarrow{d} N(0, I(\theta^*)^{-1}) \quad (n \rightarrow \infty)$$

が成り立つことから

$$\sqrt{n}(\hat{\theta}_K - \theta^*)^T I(\theta^*) \sqrt{n}(\hat{\theta}_K - \theta^*) \xrightarrow{d} \text{自由度 } K \text{ の } \chi^2 \text{ 分布}$$

ゆえに期待値は

$$\lim_{n \rightarrow \infty} E \left[\sqrt{n}(\hat{\theta}_K - \theta^*)^T I(\theta^*) \sqrt{n}(\hat{\theta}_K - \theta^*) \right] = K$$

より

$$\ell_n^*(K) = E\ell^*(\hat{\theta}_K) \rightarrow \ell^*(\theta^*) - \frac{K}{2}, \quad (n \rightarrow \infty)$$

が成り立つ。最後に $MODEL(K)$ に対する $AIC(K)$ を定義する。上記と同様に

$$\begin{aligned} & \ell(\theta^*) \\ &= \ell(\hat{\theta}_K) - \frac{1}{2}\sqrt{n}(\hat{\theta}_K - \theta^*)^T I(\theta^*) \sqrt{n}(\hat{\theta}_K - \theta^*) + o(n) \\ & \quad (n \rightarrow \infty) \end{aligned}$$

が成り立つことから $n \rightarrow \infty$ で

$$\begin{aligned} \ell^*(\theta^*) &= \ell(\theta^*) \\ &= E\ell(\hat{\theta}_K) - E \left[\frac{1}{2}\sqrt{n}(\hat{\theta}_K - \theta^*)^T I(\theta^*) \sqrt{n}(\hat{\theta}_K - \theta^*) \right] \\ &\rightarrow E\ell(\hat{\theta}_K) - \frac{K}{2} \end{aligned}$$

一方

$$\ell_n^*(K) = E\ell^*(\hat{\theta}_K) = \ell^*(\theta^*) - \frac{K}{2} + o(1)$$

を既に示した。ただし

$$\begin{aligned} E\ell(\hat{\theta}_K) &= E \left[\max_{\theta \in \Theta_K} \ell(\theta) \right] \\ &= E \left[\max_{\theta \in \Theta_K} \sum_{i=1}^n \log f(X_i; \theta) \right] \end{aligned}$$

以上から $n \rightarrow \infty$ のとき

$$\begin{aligned} \ell_n^*(K) &= E\ell^*(\hat{\theta}_K) = \ell^*(\theta^*) - \frac{K}{2} + o(1) \\ &= E\ell(\hat{\theta}_K) - \frac{K}{2} - \frac{K}{2} + o(1) \\ &= E\ell(\hat{\theta}_K) - K + o(1) \end{aligned}$$

が成り立つ。そこで

$$AIC(K) = -2\ell(\hat{\theta}_K) + 2K$$

と定義すると $AIC(K)$ は近似的に $-2\ell_n^*(K)$, すなわち期待平均対数尤度の -2 倍の推定量となる。期待平均対数尤度 $\ell_n^*(K)$ がモデルの近似の程度を測る基準として優れていることを既に述べたが, これは真の分布がわからなければ計算できない。そこで $\ell_n^*(K)$ の推定量として母集団より抽出された標本の値から $AIC(K)$ が定義された。 $AIC(K)$ は最大対数尤度の -2 倍にパラメータ数 K の 2 倍を加えたもので, 正確には $AIC(K)$ は $-2\ell_n^*(K)$ の漸近推定量である。 $AIC(K)$ が小さいほど, 期待平均対数尤度 $\ell_n^*(K)$ が大きくなるので, より真のモデルに近いと言える。

一般に確率分布の推定において, 真の分布のモデルとなる分布のパラメータ数 K を多く見積もることで見掛け上標本から得られた分布 (経験分布) にいくらかでも近い分布を見いだすことができる。 AIC の優れているところは, その定義式 $AIC(K) = -2\ell(\hat{\theta}_K) + 2K$ から, 不必要にパラメータ数 K を大きくすると $AIC(K)$ の値を大きくすることになり, 「パラメータ数を多くすると見掛けの精度が上がる」誤りを修正できることである。

6. 結言

1920 年代に, スコアの分散, 内在精度と

して定義され、統計学の論文の中に散在してはいても、式の変形などで無意識に使われていたにすぎなかった、 F -情報量が、最尤推定量の種々のすぐれた性質と相まって、現代の情報量推定法の理論的根拠となっていることをみた。また、ここではふれなかったが、 F -情報行列を計量とする Riemann 空間は負の定曲率であることがわかる。 F -情報行列が情報幾何学の出発点となったとみることもできる。

かつて、遺伝学の分野で、メンデルの遺伝学とダーウィンの自然選択論は何ら矛盾するものでなく補完的な関係にあることを統計学を使って喝破したフィッシャーのロザムステッド農事試験場でのいわば現場から生まれたであろう F -情報量の概念は、まさに湧き出る発想の泉の役割を果たした射程の長い概念であることの一端を明らかにできたと思う。

(経済学部教授)

(武蔵工業大学工学部教授)

参 考 文 献

- [1] R. A. Fisher; *On the mathematical foundations of theoretical statistics*, Phil. Trans, Roy. Soc. A, 222, 1921
- [2] C. R. Rao; On the distance between two populations, *The Indian Journal of Statistics*, 9, 1941
- [3] C. R. Rao; Information and the accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math., Soc. 37, 1945
- [4] C. R. ラオ著, 奥野忠一・長田洋・篠崎信雄・広崎昭太・古川陽子・矢島敬二・鷺尾泰俊訳; 統計的推測とその応用, 東京図書, 1977
- [5] 北川敏男; 統計情報論 I, 共立出版, 1987
- [6] 北川敏男; “フィッシャー” 「100 人の数学者」, 日本評論社, 1989
- [7] 坂元慶行・石黒真木夫・北川源四郎; 情報量統計学, 共立出版, 1983
- [8] 赤池弘次; “情報量規準 AIC とは何か”, 数理科学, No.153, 1976
- [9] 赤池弘次; “統計的検定の新しい考え方”, 数理科学, No.198, 1979
- [10] 国沢清典; 情報理論 I, 共立出版, 1983
- [11] 竹中淑子; “フィッシャー=ネイマン論争” 「数学からの 7 つのトピックス」, 培風館, 2005