

第 1 章

統計的推測法の基礎 (II)

— 尤度推測 aa —

前章では、統計的推測のうち、点推定、区間推定、そして仮説検定のそれぞれの方法における基礎概念について述べた。そこで利用した例題はいずれも単純であったため、推定や検定に必要とされる式の導出が解析的に可能であった。しかしながら、モデルの構造が複雑になると、つねに解析的に答えを出すところまでいかない場合が多い。むしろ、推定法や検定法の原理原則を提示することによって、問題解決の見通しをつけやすくなる。このような原理原則の一つが「尤度推測」という考え方で、この枠組みのおかげで推定、検定、モデルの検証と比較などが一貫して行えるようになった。

尤度の考え方の基礎は、観測データと統計モデルの組み合わせから、パラメータの値としての「尤もらしさ」を相対的に評価するところにある。例えば次のような例を考えてみよう。2項分布 $Bin(N, p)$ では、2つのパラメータ N, p がこの確率分布の性質を決定している。確率分布の性質がパラメータによってコントロールされているということは、その分布にしたがう確率変数の実現値もまたパラメータの影響を受けることになる。このことは逆に、この実現値から得られる観測情報は、パラメータに関する情報を与えることを意味し、それゆえパラメータの推測を行うことができる。もし、真の確率分布が2項分布 $Bin(10, 0.2)$ にしたがうとし、そこからサンプルサイズ 20 の標本 Y_1, Y_2, \dots, Y_{20} を得るとする。このサンプルに基づき真の分布を推測する場合、試行回数 $N = 10$ は既知であるが、 p のみが未知パラメータとなる。標本の観測値 y_1, y_2, \dots, y_{20} は前節最後の R による例でも示したような小さな値をとる傾向がある一方で、 $Y = 5$ ないし 6 以上の大きな値はほとんどとらないだろう。逆に、そのような値をとることが多いと、 $p = 0.2$ という仮定がおかしく、もう少し大きな値の方が尤もらしいのではないかと考えて不思議ではない。実はこれが観測データから得られたパラメータに関する情報の一端であり、これをフォーマルに定式化した推測の枠組みが尤度推測法である。

本章では、観測データを用いてパラメータや統計モデルに関する統計的推測法の一つの大きな枠組である尤度推測法について、その基盤となる考え方と後の章で必要となる重要な定理などを中心に述べるとともに、R を用いてどのように推定とその不確実性を評価し

ていくか、そのプロセスの概略についても述べる。

1.1 最尤原理

ここでは、統計的推測において極めて汎用的な原理である最尤原理と、その原理から導かれる推定量について解説を行う。そこで、最尤原理の理解のために、まずは区画調査の例からはじめよう。

1.1.1 最尤原理の考え方～区画調査を例に～

ここでは話を簡略化する為に、サンプルサイズを $n = 2$ とし、確率変数 $Y = (Y_1, Y_2)$ の実現値として $y = (7, 13)$ が得られたとする。さらに仮に、 λ は $\lambda = 5, 10, 15$ のいずれかの値しかとらないとする。このとき、それぞれのパラメータの値の下で $y = (7, 13)$ という値が出る確率を求めると

$$\begin{aligned} \lambda = 5 \text{ のとき } \quad P(Y = (7, 13)) &= e^{-5} \frac{5^7}{7!} \times e^{-5} \frac{5^{13}}{13!} = 0.00014 \\ \lambda = 10 \text{ のとき } \quad P(Y = (7, 13)) &= e^{-10} \frac{10^7}{7!} \times e^{-10} \frac{10^{13}}{13!} = 0.00656 \\ \lambda = 15 \text{ のとき } \quad P(Y = (7, 13)) &= e^{-15} \frac{15^7}{7!} \times e^{-15} \frac{15^{13}}{13!} = 0.00099 \end{aligned}$$

となる。これより、 $y = (7, 13)$ という実現値が最も出やすいのは、 $\lambda = 10$ のときということがわかる。すなわち、 $y = (7, 13)$ という実現値が得られたことは事実であり、これを変えることはできない。そして、その事実を最も起こりやすくするパラメータが $\lambda = 10$ となる。そうすると、一つの帰結として、この3つのパラメータ値の候補の中では $\lambda = 10$ が最も尤もらしい値と考えることができる。

このような考え方をもう少し一般化してみよう。 n 個の観測値 $Y = (Y_1, \dots, Y_n)$ の同時確率分布は

$$P(Y = y) = \prod_{i=1}^n P(Y_i = y_i) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} \quad (1.1)$$

となる。これを λ の関数とみなし、 $L(\lambda)$ とおく。この関数の値を大きくすることは、それだけ観測値が出現しやすいパラメータ値を求めていることになる。そこで、 λ を $0 < \lambda < \infty$ の範囲で連続的に変化させて、 $L(\lambda)$ を最大にする λ を求めれば、先程のような最も尤もらしい λ の値を導くことができる。このように、確率分布が記述できているとき、これをパラメータの関数とみなし、その関数を最大化することによりパラメータの値を推定する方法を最尤法 (maximum likelihood method) という。

ところで、対数をとっても大小関係は変わらないから、最大となるパラメータ値を求める際、対数尤度関数を用いることが多い。これは解析的に解く場合に計算が楽になることに加え、解析的に解けない場合でも確率あるいは確率密度を掛け合わせて小さな値をとる

関数を扱うよりも、対数を取って和の計算をする方が数値的に安定すること、そして対数尤度関数を計算することのメリットが数多くあるからである。

$$\log L(\lambda) = \sum_{i=1}^n \log(y_i!) - n\lambda + \sum_{i=1}^n y_i \log \lambda \quad (1.2)$$

この例の場合、 λ に関して微分し 0 とおいた方程式を解くことで、 $\hat{\lambda}(y) = \bar{y}$ という明示的な解を得ることができる。この解の形自体は実現値 y の値に依存しないので、 $\hat{\lambda}(Y) = \bar{Y}$ として確率変数を用いて一般的に書くことができる。

最尤法は観測値の不確実性を確率分布を通して考慮に入れており、また直感的にも受け入れ易い推定法である。最尤法の原理は、この例に限らず一般的な確率モデルにおいても利用できる非常に守備範囲の広い方法であり、しかも後に述べるように幾つかの望ましい性質をもつ推定方法である。

1.1.2 最尤推定の一般的定式化

それでは、最尤推定法を一般的に定式化してみよう。

定義 1.1 [尤度関数] 確率変数 $Y = (Y_1, \dots, Y_n)$ の確率関数あるいは確率密度関数を $f(y_1, y_2, \dots, y_n; \theta)$ とする。 $\theta = (\theta_1, \dots, \theta_d)^t$ は推測の対象となるパラメータで、 $\theta \in \Theta \subseteq R^d$ とする。観測値 $Y = y$ を固定し、 $L(\theta) = f(y|\theta)$ を未知パラメータ $\theta \in \Theta$ の関数とみなすとき、この関数

$$L(\theta) = f(y_1, y_2, \dots, y_n; \theta) \quad (1.3)$$

を尤度関数 (likelihood function) とよぶ。また、尤度関数の対数 $\log L(\theta)$ を対数尤度関数 (log-likelihood function) とよぶ。

定義 1.2 [最尤推定量] 観測値 $Y = y$ が与えられたとき、尤度関数 $L(\theta)$ を最大にするパラメータ値 $\theta(y)$ を最尤推定値 (maximum likelihood estimate) とよぶ。また、これを確率変数 Y の関数とみなしその統計的な性質を評価したいとき、 $\theta(Y)$ を最尤推定量 (maximum likelihood estimator) とよぶ。

先にも述べたが、最尤法によるパラメータの推定値は対数尤度関数

$$l(\theta) = \log L(\theta) = \log f(y_1, y_2, \dots, y_n; \theta) \quad (1.4)$$

を最大にする θ を求めることによって求められることが多い。そこで、とくに対数尤度関数の導関数

$$u(\theta) = \frac{\partial}{\partial \theta} \log L(\theta) = \frac{\partial}{\partial \theta} \log f(y_1, y_2, \dots, y_n; \theta) \quad (1.5)$$

をスコア関数 (score function) とよぶ。最尤推定量を $\hat{\theta}(Y)$ とおくと、 $\hat{\theta} = \hat{\theta}(Y)$ が Θ の内点 (端点ではないこと) であれば、最尤推定量はスコア方程式

$$u(\theta) = 0 \quad (1.6)$$

の解でもある。

なおここでは、 θ がベクトルである場合も想定しているので、実際にはスコア方程式は以下の同時方程式となる。

$$\frac{\partial}{\partial \theta} L(\theta) = 0 \iff \begin{cases} \frac{\partial}{\partial \theta_1} \log L(\theta) = 0 \\ \dots \\ \frac{\partial}{\partial \theta_d} \log L(\theta) = 0 \end{cases} \quad (1.7)$$

1.1.3 最尤推定法の例題：解析的な解法が可能な場合

この節では、解析的に解ける簡単な例から最尤推定値の実際の計算を行うことにする。

例 1.1 [2 項分布における p の推定] 確率変数 Y_1, Y_2, \dots, Y_n が独立同一に 2 項分布 $Bin(N, p)$ (N は既知) に従うとする。このとき、 $Y = (Y_1, Y_2, \dots, Y_n)$ に基づく p の最尤推定量 $\hat{p}(Y)$ を求め、 $\hat{p}(Y)$ の期待値と分散を計算したい。いま、確率変数 Y_1, \dots, Y_n の同時分布は

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^n \binom{N}{y_i} p^{y_i} (1-p)^{N-y_i}. \quad (1.8)$$

で表せるから、パラメータ p 対数尤度関数は、

$$\log L(p) = \log \prod_{i=1}^n \binom{N}{y_i} + \left(\sum_{i=1}^n y_i \right) \log p + \left(\sum_{i=1}^n (N - y_i) \right) \log(1-p). \quad (1.9)$$

となり、これを p で偏微分して以下のようにスコア関数が求められる。

$$\frac{\partial}{\partial p} l(p) = \left(\sum_{i=1}^n y_i \right) \frac{1}{p} - \left(\sum_{i=1}^n (N - y_i) \right) \frac{1}{1-p} \quad (1.10)$$

スコア方程式を解くと

$$\hat{p}(Y) = \frac{1}{Nn} \sum_{i=1}^n y_i \quad (1.11)$$

を得る。上のようにして得られた推定量もまた確率変数であるから、確率変数の独立性などに注意して $\hat{p}(Y)$ の期待値と分散を計算すると

$$\begin{aligned} E[\hat{p}(Y)] &= E \left[\frac{1}{Nn} \sum_{i=1}^n Y_i \right] = \frac{1}{Nn} E \left[\sum_{i=1}^n Y_i \right] = \frac{n}{Nn} E[Y_1] = \frac{1}{N} Np = p, \\ V[\hat{p}(Y)] &= V \left[\frac{1}{Nn} \sum_{i=1}^n Y_i \right] = \left(\frac{1}{Nn} \right)^2 V \left[\sum_{i=1}^n Y_i \right] \\ &= \frac{n}{N^2 n^2} n V[Y_1] = \frac{1}{nN^2} Np(1-p) = \frac{1}{Nn} p(1-p). \end{aligned} \quad (1.12)$$

となり、 $\hat{p}(Y)$ は不偏推定量で、サンプル数 n または 1 回の試行数 N が大きくなると分散が小さくなることが分かる。

例 1.2 [相対漁獲効率の推定] 種類の異なる 2 つの漁具を 1 時間設置し、得られた漁獲尾数の情報を基に、それらの漁獲効率の違いを調べる。漁具 1 を 1 時間設置した際の漁獲尾数の期待値を λ 、漁具 2 のそれを $\phi\lambda$ とする。ここで、 ϕ は相対漁獲効率を示すパラメータである。漁具 1, 2 における漁獲尾数を Y_1, Y_2 とする。漁獲尾数はそれぞれポアソン分布に従うとすると、 λ および ϕ の最尤推定量を求めたい。まず、パラメータ λ, ϕ に対する尤度関数および対数尤度関数は

$$\begin{aligned} L(\lambda, \phi) &= e^{-\lambda} \frac{\lambda^{y_1}}{y_1!} \times e^{-\phi\lambda} \frac{(\phi\lambda)^{y_2}}{y_2!} = e^{-(1+\phi)\lambda} \frac{\lambda^{y_1+y_2} \phi^{y_2}}{y_1! y_2!} \\ l(\lambda, \phi) &= \log L(\lambda, \phi) = -(1+\phi)\lambda + (y_1 + y_2) \log \lambda + y_2 \log \phi - \log(y_1! y_2!) \end{aligned} \quad (1.13)$$

となる。この対数尤度関数を λ, ϕ で偏微分し、以下の連立スコア方程式を得る。

$$\begin{cases} \frac{\partial}{\partial \lambda} l(\lambda, \phi) = -(1+\phi) + \frac{y_1 + y_2}{\lambda} = 0 \\ \frac{\partial}{\partial \phi} l(\lambda, \phi) = -\lambda + \frac{y_2}{\phi} = 0 \end{cases} \quad (1.14)$$

これを解けば最尤推定値の組み合わせ

$$\hat{\lambda} = y_1, \quad \hat{\phi} = \frac{y_2}{y_1} \quad (1.15)$$

を得る。(以下、推定量が依存している確率変数が明らかな場合、それを省略することもある。)

例 1.3 [対立遺伝子頻度の推定] ある魚種の繁殖集団から n 個体のサンプリングを行い、核 DNA のある遺伝子座における対立遺伝子を観察した。この結果、 n 個体において 3 種類の対立遺伝子を観測した。また、それぞれの対立遺伝子の観測数をベクトル (Y_1, Y_2, Y_3) と表す。ここで、 $\sum_{j=1}^3 Y_j = 2n$ である。真の対立遺伝子頻度をベクトル $p = (p_1, p_2, p_3)$ (ただし $p_1 + p_2 + p_3 = 1$) で表し、確率変数 (Y_1, Y_2, Y_3) が多項分布 $Multi(2n, p)$ に従うとすると、 p_1, p_2, p_3 の最尤推定量を求めたい。多項分布 $Multi(2n, p)$ の確率分布は

$$Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = 2n - y_1 - y_2) = \frac{(2n)!}{y_1! y_2! (2n - y_1 - y_2)!} p_1^{y_1} p_2^{y_2} p_3^{2n - y_1 - y_2} \quad (1.16)$$

である。微分すればと 0 となる定数項を無視すると、対数尤度関数は

$$\log L(p_1, p_2) = y_1 \log p_1 + y_2 \log p_2 + (2n - y_1 - y_2) \log(1 - p_1 - p_2) \quad (1.17)$$

となる。 $p_1 + p_2 + p_3 = 1$ より自由パラメータは 2 つに限られることに注意すると、同時スコア方程式が

$$\begin{aligned} \frac{\partial}{\partial p_1} l(p_1, p_2) &= \frac{y_1}{p_1} - \frac{2n - y_1 - y_2}{1 - p_1 - p_2} = 0 \\ \frac{\partial}{\partial p_2} l(p_1, p_2) &= \frac{y_2}{p_2} - \frac{2n - y_1 - y_2}{1 - p_1 - p_2} = 0 \end{aligned} \quad (1.18)$$

のように得られ、下記のような自然な推定値が得られる。

$$\hat{p}_1 = \frac{y_1}{2n}, \quad \hat{p}_2 = \frac{y_2}{2n}, \quad \hat{p}_3 = \frac{y_3}{2n}, \quad (1.19)$$

例 1.4 [指数分布のパラメータの推定] ある魚種の産卵から死亡までの時間標本 Y_1, Y_2, \dots, Y_n を観測した。この確率変数に指数分布

$$f(y|\theta) = \frac{1}{\theta} \exp\left\{-\frac{y}{\theta}\right\} \quad (y > 0, \theta > 0) \quad (1.20)$$

を仮定するとき、 θ の最尤推定量を求めたい。これまでと同様に、尤度関数と対数尤度関数がそれぞれ

$$L(\theta) = \prod_{i=1}^n f(y_i|\theta) = \frac{1}{\theta^n} \exp\left\{-\frac{\sum_{i=1}^n y_i}{\theta}\right\} \quad (1.21)$$

$$l(\theta) = \log L(\theta) = -n \log \theta - \frac{\sum_{i=1}^n y_i}{\theta}$$

のように表現できるから、スコア方程式を解くことで以下のように推定値が得られる。

$$\frac{\partial}{\partial \theta} l(\theta) = -n/\theta + \frac{\sum_{i=1}^n y_i}{\theta^2} = 0 \iff \hat{\theta} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y} \quad (1.22)$$

練習問題 1.1 [クジラの潜水浮上] ある鯨種の潜水浮上プロセスを観測した。2 頭のクジラ (個体 A と B) に追尾可能なタグを装着し、それぞれの潜水時間 (浮上と浮上の間隔) Y_{A1}, \dots, Y_{A,n_A} と Y_{B1}, \dots, Y_{B,n_B} を記録した。それぞれの個体において、毎回の潜水時間は独立かつ同一に指数分布

$$f(x_{Ai}; \sigma_A) = \frac{1}{\sigma_A} e^{-x_{Ai}/\sigma_A} \quad \text{for } x_{Ai} > 0, \sigma_A > 0 \quad (i = 1, 2, \dots, n_A)$$

$$f(x_{Bi}; \sigma_B) = \frac{1}{\sigma_B} e^{-x_{Bi}/\sigma_B} \quad \text{for } x_{Bi} > 0, \sigma_B > 0 \quad (i = 1, 2, \dots, n_B)$$

に従うとする。このとき、個体 A と個体 B の平均潜水時間 σ_A, σ_B の最尤推定量とその標準偏差を求めよ。さらに、以下の観測値が得られたとき、それぞれの推定値を求めよ。

個体 A 0.16 6.96 23.03 6.89 0.55 10.39 6.79 2.91 0.62 10.33

 3.64 3.04 2.66 1.45 1.44 2.85 1.71 3.22 0.77 5.53

個体 B 0.70 0.72 1.70 2.64 4.70 1.13 0.97 2.39 0.85 0.47

 8.86 4.34 5.19 13.18 3.97 0.30 5.60 11.84

また、 $\sigma = \sigma_A = \sigma_B$ と仮定したときの σ の最尤推定量も求めよ。

練習問題 1.2 [分散既知の場合の正規分布の平均値の推定] 確率変数 Y_1, Y_2, \dots, Y_n が独立同一に正規分布 $N(\mu, \sigma^2)$ (σ は既知) に従うとする。このとき、 $Y = (Y_1, Y_2, \dots, Y_n)$ に基づく μ の最尤推定量 $\hat{\mu}(Y)$ を求めよ。また、 $\hat{\mu}(Y)$ の期待値と分散を計算せよ。

練習問題 1.3 [分散未知の場合の正規分布の平均値の推定] 確率変数 Y_1, Y_2, \dots, Y_n が独立同一に正規分布 $N(\mu, \sigma^2)$ に従うとする. このとき, $Y = (Y_1, Y_2, \dots, Y_n)$ に基づく μ, σ^2 の最尤推定量 $\hat{\mu}(Y), \hat{\sigma}^2(Y)$ を求めよ. また, $\hat{\sigma}^2(Y)$ の期待値を計算せよ.

例 1.5 [正規線形回帰モデルにおけるパラメータの推定] 線形回帰モデルは後の章でも取り上げるが, ここでは最尤法の中に最小 2 乗法が現れることを, 1 次元の回帰分析を通してみていく. いま以下のような線形回帰モデルを考える. X_i が説明変数で, それに依存して変動する Y_i が応答変数である.

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \dots, n. \quad (1.23)$$

これは, $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$ のようにも表せる. このとき, α, β, σ^2 が未知パラメータであり, これらの尤度関数と対数尤度関数が以下のように得られる.

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right\} \\ l(\alpha, \beta, \sigma^2) &= \log L(\alpha, \beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \end{aligned} \quad (1.24)$$

これにより, 対数尤度関数を最大にする α, β は最小 2 乗法による解と同値であることが分かる. 実際, 対数尤度関数 $l(\alpha, \beta, \sigma^2)$ を偏微分した連立方程式

$$\begin{aligned} \frac{\partial}{\partial \alpha} l(\alpha, \beta, \sigma^2) &= -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-1) = 0 \\ \frac{\partial}{\partial \beta} l(\alpha, \beta, \sigma^2) &= -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-x_i) = 0 \\ \frac{\partial}{\partial \sigma^2} l(\alpha, \beta, \sigma^2) &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = 0 \end{aligned} \quad (1.25)$$

を解けば, 3 つパラメータの最尤推定値が得られる.

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}, \\ \hat{\beta} &= \frac{\sum_{i=1}^n (y_i - \bar{y})x_i}{\sum_{i=1}^n (x_i - \bar{x})x_i} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2. \end{aligned} \quad (1.26)$$

一般に $\hat{\theta}$ は明示的に解くことができない場合が多く, したがって準 Newton 法などによる解法が標準的である. また, Scoring 法 (Fisher's method of scoring (Osborn

(1992)) という統計学特有の最適化法もある。この Scoring 法は、Newton 法におけるヘッシアン行列をその期待値、すなわち Fisher 情報行列 (後述) で置き換える方法で、次節で述べる Fisher 情報行列が正値行列であるから必ず探索方向は尤度の値を大きくする方向へ更新される。また収束した段階で Fisher 情報行列の逆行列 $I^{-1}(\hat{\theta})$ が得られるため、 $\hat{\theta}$ の分散共分散行列の推定値 (後述) が副産物として同時に得られる。

[一口メモ] 先にも述べたように、直感的な意味での最尤原理は「観測値の出現確率を最大にするような、最も尤らしいパラメータを推定値とする」である。通常理解としてこれで十分であるが、数学的には、最尤推定は観測値の経験分布とモデル $M = \{f(y; \theta); \theta \in \Theta\}$ との K-L 擬距離 (Kullback-Leibler distance) を最小にするようなパラメータを求めている。すなわち幾何学的には、経験分布からモデルへの射影先が $f(y; \hat{\theta})$ となる。そのような意味で、直感的な意味に加えて数学的な裏づけも与えられていることを注意しておく。

1.2 最尤推定量の不確実性の評価：最尤推定量の漸近理論

尤度推測における長所の一つに、大標本理論、すなわち漸近的な性質が一般的性質として確立している点が挙げられる。ここでは、最尤推定量と尤度比統計量の漸近的性質を紹介する。

1.2.1 最尤推定量の漸近的性質

本題に入る前に、確率論および統計学における非常に重要な定理を2つ紹介する。ただし、証明と詳細な説明は省略する。たとえば Lehmann (20xx), 稲垣 (19xx), ... などの成書をご覧ください。

定理 1.1 大数の (弱) 法則 確率変数 X_1, X_2, \dots, X_n は独立かつ同一の確率分布に従い、その分布の平均値を μ 、分散は σ^2 とする。また、標本平均を $\bar{X}_n = \sum_{i=1}^n X_i/n$ とおく。このとき、任意の正の数 $\varepsilon > 0$ に対して

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1 \quad (1.27)$$

が成り立つ。このように、標本平均 \bar{X}_n が分布の平均値 μ に収束することを、 \bar{X}_n が μ に確率収束 (convergence in probability) するといひ、

$$\bar{X}_n \rightarrow \mu \text{ in prob} \quad (1.28)$$

などとも記述する。 ■

定理 1.2 中心極限定理 確率変数 X_1, X_2, \dots, X_n は独立かつ同一の確率分布に従い、その確率分布の平均値を μ 、分散は σ^2 とする。このとき、標本平均を標準化した統計量

$$Z_n = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \quad (1.29)$$

に対して、

$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad (1.30)$$

が成り立つ。このことを、 Z_n は標準正規分布 $N(0, 1)$ に分布収束 (convergence in distribution) する、といい、

$$Z_n = \sqrt{n} \frac{\bar{Y}_n - \mu}{\sigma} \rightarrow N(0, 1) \quad (1.31)$$

などとも記述する。 ■

大雑把に言うと、 n が大きくなると確率変数列 X_n の標本平均はその期待値に近づき、その近づく様子も \sqrt{n} の拡大眼鏡で見えていくと Z_n が正規分布に従いながらその期待値 0 に近づいていくことを示している。

図を入れる

次に最尤推定量の漸近的性質を 2 つ挙げるが、ここでは少しでも話をわかりやすくするために、確率変数が独立かつ同一に確率分布 $f(y_i; \theta)$ に従うと仮定し、 d 次元のパラメータ θ について最尤推定を行うことを想定する。この仮定が成り立たないときも、特別な事情がない限り、これから述べる最尤推定量の性質は成り立つ。また独立同一性の下では、対数尤度関数とスコア関数がそれぞれ

$$l(\theta) = \log L(\theta) = \log \prod_{i=1}^n f(y_i; \theta) = \sum_{i=1}^n \log f(y_i; \theta) \quad (1.32)$$

$$u(\theta) = \sum_{i=1}^n u_i(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(y_i; \theta) \quad (1.33)$$

のように書ける。したがって、 $X_i = \log f(y_i; \theta)$ や $X_i = u_i(\theta)$ とおけば、これらに対しても大数の法則などの定理が使えることになる。なお、これまで同様に θ は d 次元 (縦) ベクトルとし、以下に示すように θ に関する 1 階微分はベクトルに、2 階微分は行列をもたらすことに注意する。

$$\frac{\partial}{\partial \theta} = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_d} \end{pmatrix}, \quad \frac{\partial^2}{\partial \theta \partial \theta^t} = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_d} \\ \vdots & & \vdots \\ \frac{\partial^2}{\partial \theta_d \partial \theta_1} & \cdots & \frac{\partial^2}{\partial \theta_d^2} \end{pmatrix} \quad (1.34)$$

定理 1.3 最尤推定量の漸近的性質 適当な正則条件の下で、標本数 n が十分大きい時、最尤推定量 $\hat{\theta}$ は真のパラメータ θ に確率収束し、 $\sqrt{n}(\hat{\theta} - \theta)$ の分布は正規分布に近づく。すなわち、

$$\begin{aligned}\hat{\theta} &\xrightarrow{P} \theta && \text{(一貫性)} \\ \sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{D} N(0, I^{-1}(\theta)) && \text{(漸近正規性)}\end{aligned}\quad (1.35)$$

が成立する。ただし、 $I(\theta)$ は 1 標本あたりの Fisher 情報行列

$$I(\theta) = E_{\theta} \left[\left(\frac{\partial}{\partial \theta} \log f(y_i; \theta) \right) \left(\frac{\partial}{\partial \theta^t} \log f(y_i; \theta) \right) \right] = E_{\theta} \left[- \frac{\partial^2}{\partial \theta \partial \theta^t} \log f(y_i; \theta) \right] \quad (1.36)$$

である。 ■

式 (1.36) の等号は一見難しく見えるが、導出は難しくない。簡略のために $d = 1$ とし、連続型分布で記述すると、確率密度の積分が 1 であり、この場合に微分と積分の交換可能性から、

$$\begin{aligned}\int f(y_i; \theta) dy &= 1 \text{ だ} \iff \frac{\partial}{\partial \theta} \int f(y_i; \theta) dy = 0 \iff \int \frac{\partial}{\partial \theta} f(y_i; \theta) dy = 0 \\ \therefore \int \left(\frac{\partial}{\partial \theta} \log f(y_i; \theta) \right) f(y_i; \theta) dy &= E \left[\frac{\partial}{\partial \theta} \log f(y_i; \theta) \right] = E[u_i(\theta)] = 0\end{aligned}\quad (1.37)$$

となり、さらにもう一度微分すると、

$$\begin{aligned}\frac{\partial}{\partial \theta} \int \left(\frac{\partial}{\partial \theta} \log f(y_i; \theta) \right) f(y_i; \theta) dy &= 0 \\ \iff \int \left(\frac{\partial^2}{\partial \theta^2} \log f(y_i; \theta) \right) f(y_i; \theta) dy + \int \left(\frac{\partial}{\partial \theta} \log f(y_i; \theta) \right)^2 f(y_i; \theta) dy &= 0 \\ \therefore E \left[\frac{\partial^2}{\partial \theta^2} \log f(y_i; \theta) \right] &= -E \left[\left(\frac{\partial}{\partial \theta} \log f(y_i; \theta) \right)^2 \right]\end{aligned}\quad (1.38)$$

となる。離散型の場合も同様で、 $d > 1$ 場合には微分がベクトルや行列を含むだけで、形式的には導出は全く同じである。

また厳密な評価が必要となる場面以外、実用の場面で式 (1.36) の計算を行うことはほとんどない。これは、大数の法則から

$$-\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta^t} \log f(y_i; \theta) \xrightarrow{P} I(\theta) \quad (1.39)$$

が成り立ち、加えて左辺が Newton 法の数値計算で内部的に計算されるため、実質的には期待値を取らない観測された **Fisher 情報行列** で代用することが多いからである。

最尤推定量は一般に不偏推定量とは限らないが、一貫性が成り立ち漸近的には不偏となる。厳密には、この一貫性は正しいモデルが選ばれている場合に成り立つ性質であるが、後述する一般化線形モデルのように、分散の構造は間違えても期待値に関する構造を間違えなければ少なくとも一貫性は保証されるという場合もある。

また、最尤推定量に対する漸近正規性のおかげで、パラメータの標準誤差の評価や信頼区間の構成にも利用できる。この説明をするためには、上記の式 (1.35) の漸近正規性が

$$\hat{\theta} \xrightarrow{D} N(\theta, I_n^{-1}(\theta)) \quad (1.40)$$

と書き直せること注意する。ここで $I_n(\theta)$ は n 個の観測値に対する Fisher 情報行列で、

$$I_n(\theta) = E_{\theta} \left[\left(\frac{\partial}{\partial \theta} \log f(y; \theta) \right) \left(\frac{\partial}{\partial \theta^t} \log f(y; \theta) \right) \right] = E_{\theta} \left[-\frac{\partial^2}{\partial \theta \partial \theta^t} \log f(y; \theta) \right] \quad (1.41)$$

であり、確率変数が独立同一の場合には特別に

$$I_n(\theta) = E_{\theta} \left[\sum_{i=1}^n -\frac{\partial^2}{\partial \theta \partial \theta^t} \log f(y_i; \theta) \right] = \sum_{i=1}^n E_{\theta} \left[-\frac{\partial^2}{\partial \theta \partial \theta^t} \log f(y_i; \theta) \right] = nI(\theta) \quad (1.42)$$

が成り立ち、式 (1.35) となる。

式 (1.40) より最尤推定量 $\hat{\theta}$ の分散共分散行列が近似的に

$$V[\hat{\theta}] \approx I_n^{-1}(\theta) \quad (1.43)$$

と評価できるから、この関係を利用して最尤推定値 $\hat{\theta}$ の分散共分散行列を

$$\widehat{V}[\hat{\theta}] \approx \frac{1}{n} I^{-1}(\hat{\theta}) \quad (1.44)$$

として推定できる。あるパラメータ θ_d の推定値 $\hat{\theta}_d$ の精度を表すために推定値の標準偏差を推定値に付与することが多い。この値を標準誤差という。なお詳細は省くが、最尤推定量は、推定量の分散の下限を漸近的に達成する漸近有効推定量となっている。その意味では、最も精度の高い推定量でもある。(ただし、利用しているモデルが正しいことが前提である)

例 1.6 [最尤推定量の挙動に関するシミュレーション] 2 項分布 $Bin(n, p)$ に従う確率変数 Y を観測してパラメータ p の推定を行う。 p の最尤推定量 $\hat{p} = Y/n$ の確率分布をシミュレートしたい。仮に、真のパラメータの値を $p = p_0$ とする。 $n = 20, 50, 100, 1000$ をそれぞれ仮定したとき、 \hat{p} のヒストグラムを 10000 回の繰り返し (すなわち、各 n の値の下で 10000 個の \hat{p} を求め)、そのヒストグラムを書いてみるとヒストグラムがどんどんつぶれていくが、 $\sqrt{n}(\hat{p} - p_0)/\sqrt{p_0(1 - p_0)}$ のように拡大眼鏡で見るとヒストグラムが標準正規分布 $N(0, 1)$ に近づくことが分かる。

```

1
2 p0 <- runif(1,0,1) # p0 の値を一様分布 U(0,1) に従う乱数をひとつ生成して決定
3 p0
4 Nvector <- c(20, 50, 100, 1000) # 想定するサンプル数 N
5 Iteration <- 10000 # シミュレーションの繰り返し数
6 phat <- list(NULL) # p の推定値を保存するために用意

```

```

7  par(mfrow=c(2,2))    # グラフの画面を 4 分割するコマンド
8
9  for(i in 1:length(Nvector) )    # n=20, 50, 100, 1000 を変えて計算させるループ
10 {
11  phat[[i]] <- numeric(Niteration)# n=20, 50, 100, 1000 の場合それぞれで違う箱に入れる
12  for(j in 1:Niteration)    # 10000 回の繰り返しのループ
13  {
14  y <- rbinom(1, Nvector[i], p0) # データ Y を作成するために 2 項分布に従う乱数をひとつ生成
15  phat[[i]][j] <- y/Nvector[i]   # p の推定量
16  }
17  # ヒストグラムの作成
18  hist(phant[[i]], main=paste("N=", Nvector[i], ", p0=", round(p0,4) ),
19  xlab="Estimate of p", freq=F, xlim=c(0,1), col="grey" )
20  # 真の値をプロット. pch=4 は×印, cex=1.5 は×の大きさ指定
21  points( p0,0, col="red", cex=1.5, pch=4)
22 }
23
24 win.graph()
25 par(mfrow=c(2,2))
26 for(i in 1:length(Nvector) )
27 {
28  # ヒストグラムの作成
29  hist(sqrt(Nvector[i])*(phant[[i]]-p0)/sqrt(p0*(1-p0)),
30  main=paste("N=", Nvector[i], ", p0=", round(p0,4) ),
31  xlab="SQRT(N) (phant-p0)/sqrt(p0*(1-p0))", freq=F, xlim=c(-4,4), col="grey" )
32  curve( dnorm(x, 0, 1), add=T) # ヒストグラムに正規分布の密度関数を重ね書き
33 }

```

(Ch2.sim1.r)

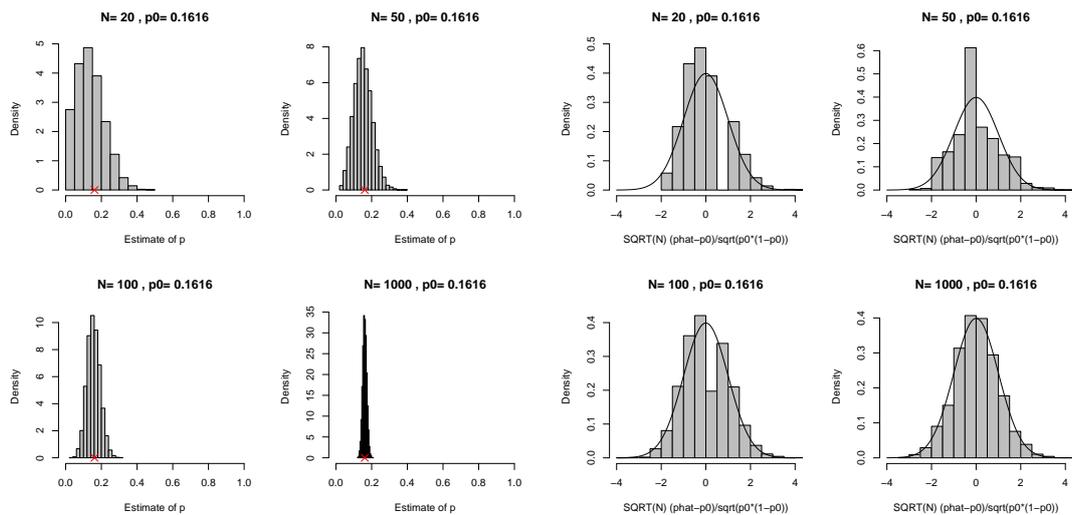


図 1.1: \hat{p} (左) と $\sqrt{N}(\hat{p} - p_0)/\sqrt{p_0(1 - p_0)}$ (右) のヒストグラム

1.2.2 パラメータの変換への対応

次に、以下のようなパラメータの変換を考える。これは、例えば正規分布 $N(\mu, \sigma^2)$ のパラメータを推定した後に変動係数 σ/μ の推定も考える、というような場合である。

$$g: \Theta \rightarrow R \quad (1.45)$$

このようなパラメータの変換をした場合、 $\phi = g(\theta)$ に対する推定量や誤差の評価をどのようにすればよいだろうか。あるいは、 θ の推測の結果を利用するにはどうすればよいだろうか。推定量に対しては良く知られた性質として、 $\hat{\phi} = g(\hat{\theta})$ が ϕ の最尤推定量となる。また、最尤推定量 $\hat{\phi}$ に対して

$$\hat{\phi} - \phi \rightarrow N(0, v_\phi(\theta)) \quad \text{in dist} \quad (1.46)$$

が成立する。ただし、 $v_\phi(\theta)$ は $\hat{\phi}$ の漸近分散で、Taylor 展開を用いると、

$$v_\phi(\theta) = \frac{\partial \phi}{\partial \theta^t} I^{-1}(\theta) \frac{\partial \phi}{\partial \theta} \quad (1.47)$$

で得られる。式 (1.47) のように変換されたパラメータの分散を計算する方法をデルタ法という。

1.3 尤度推測における数値計算：簡単な実例を通して

この節では、例題を通してさらに尤度関数を用いた推測の理解を深めることを目的とする。特に、 R で尤度関数を定義し、それを最適化することでパラメータ値と標準誤差を求めることにする。

1.3.1 ポアソン分布モデルと Presence-absence モデル

いま、ある生物種の個体群密度の調査を考える。生息域内において、ある一定の面積のサイトを K 箇所ランダムに選び、そのサイト内の個体数を正確に観測する場合（ポアソン分布モデル、略して P モデル）と、サイト内に個体がいるかないかのみを観測する場合（Presence-absence モデル、略して P/A モデル）の 2 つのモデルを考える。それぞれのサイト内に生息する個体数を独立同一なポアソン分布に従う確率変数とし、

$$X_1, X_2, \dots, X_K \stackrel{iid}{\sim} Po(\lambda) \quad (1.48)$$

とおく。ポアソン分布モデルでは X_i が直接観察できるが、Presence-absence モデルでは

$$Y_i \sim Bin(1, p(\lambda)), \quad p(\lambda) = P(X_i > 0) = 1 - P(X_i = 0) = 1 - e^{-\lambda} \quad (1.49)$$

となる（この和をさらに 2 項分布として表現することもできるが、ここでは K 個のベルヌーイ確率変数とみなしておく）。

両モデルとも非常に単純なモデルで、P モデルの尤度は尤度原理の説明で述べた通り、

$$\log L_1(\lambda) = \sum_{i=1}^K \log \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) \quad (1.50)$$

であり、最尤推定量は解析的に $\hat{\lambda} = \bar{X}$ と求まる。また、P/A モデルも尤度は

$$\log L_2(\lambda) = \sum_{i=1}^K \log [p(\lambda)^{y_i} (1 - p(\lambda))^{1-y_i}] \quad (1.51)$$

と書け、 $\hat{\lambda} = -\log(1 - \bar{Y})$ が得られる。

これらの例では最尤推定量はどちらの場合も数値計算の必要がないが、この簡単な例を通して最尤推定値を数値的に求める方法を習得しよう。 $\lambda = 3, K = 50$ と仮定して、ミュレーションデータを生成し、その観測値を基に対数尤度の最適化を行う。

```
1 # Parameter setting
2 set.seed(2018)
3 lambda <- 3
4 K <- 50
5
6 # Data generation
7 x <- rpois(K, lambda)
8 y <- as.numeric(x>0)
9 t(data.frame(x,y))
```

(poisson.data.r)

```
> t(data.frame(x,y))
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
x    2    3    1    1    3    2    3    1    6    3    2    4    7    4
y    1    1    1    1    1    1    1    1    1    1    1    1    1    1
...
[,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49] [,50]
x     3     1     5     4     3     3     3     0     6     1     1     4
y     1     1     1     1     1     1     1     0     1     1     1     1
```

(poisson.data.r の実行結果)

次に 2 つのモデルの尤度関数を定義する。本来 $\lambda > 0$ という制約を入れた尤度を定義すべきであるが、それは次節に譲り、ここでは説明を簡明にするために制約なしでコーディングした。

```
1 # Definition of negative loglikelihood (P model)
2 NLL1 <- function(lambda){
3   logLL <- dpois(x, lambda, log=T)
```

```

4   obj <- (-1.0)*sum(logLL)
5   obj
6 }
7
8 # Definition of negative loglikelihood (P/A model)
9 NLL2 <- function(lambda){
10  p <- 1-exp(-lambda)
11  logLL <- dbinom(y, 1, p, log=T)
12  obj <- (-1.0)*sum(logLL)
13  obj
14 }

```

(poisson.like.r)

まずは尤度の概形を捉えると、観測において情報損失のない P モデルの方がシャープに入に対するコントラストを与えていることが分かる。このことと Fisher 情報量とはとても深い関係がある。Fisher 情報量が対数尤度の 2 階微分で計算され、対数尤度の最大値付近での曲がり具合（曲率）がそのまま情報の大小とも関係し、それが推定の精度となって反映されるのである。

```

1 # Shape
2 lambda.vec <- seq(1, 6, length.out=100)
3 NLL1.vec <- sapply(lambda.vec, NLL1)
4 NLL2.vec <- sapply(lambda.vec, NLL2)
5 Like1.prof <- NLL1.vec-min(NLL1.vec)
6 Like2.prof <- NLL2.vec-min(NLL2.vec)
7 plot(lambda.vec, Like1.prof, type="l", xlab="lambda", ylab="Likelihood profile")
8 points(lambda.vec, Like2.prof, type="l", lty=2)
9 legend(2, 50, lty=c(1,2), legend=c("P model", "P/A model"), bty="n")

```

(poisson.shape.r)

実際パラメータを推定してみると、最適化のために `optim` を使い、結果を取り出すコードが以下の通り。 `hessian=T` と指定しているので、計算後に最終的なヘッシアンが取り出せ (NLL のように負の対数尤度であらかじめ定義しているので、ヘッシアンにマイナス 1 を乗じなくてよい)、この場合にはパラメータが 1 つなのでヘッシアンも 1 次元となり、その逆数の平方根が標準誤差となる。

```

1 Res1 <- optim(2, NLL1, method="BFGS", hessian=T)
2 lambda1.est <- Res1$par
3 lambda1.se <- sqrt(1/Res1$hessian)
4
5 Res2 <- optim(2, NLL2, method="BFGS", hessian=T)
6 lambda2.est <- Res2$par
7 lambda2.se <- sqrt(1/Res2$hessian)
8

```

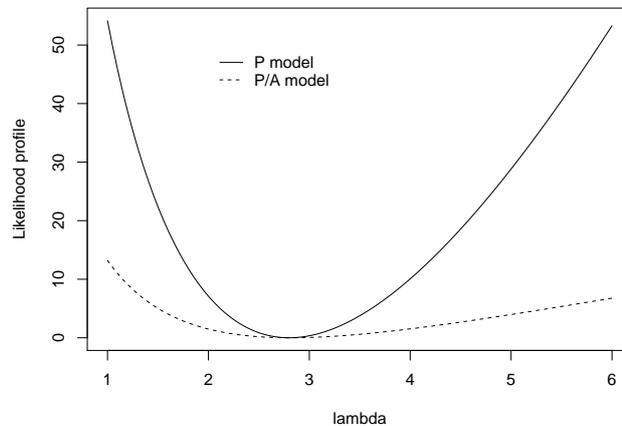


図 1.2: P モデルと P/A モデルの負の対数尤度の比較. いずれも最小値を引いているので, 0 が最小値となるように基準化.

```

9 data.frame(lambda1.est,lambda1.se, lambda2.est,lambda2.se)
10
11 points(lambda1.est, 0, pch=18, col="red")
12 points(lambda2.est, 0, pch=19, col="blue")

```

(poisson.est.r)

```

> data.frame(lambda1.est,lambda1.se, lambda2.est,lambda2.se)
  lambda1.est lambda1.se lambda2.est lambda2.se
1          2.8 0.2366432    2.813411 0.5597617

```

(poisson.est.r の実行結果)

尤度プロファイルから, より正確な信頼区間を求めることができるが, これ以降の章で別途扱うこととする. (ここに入れるか考え中)

1.3.2 occupancy モデル

それでは, P モデルと P/A モデルよりももう少し複雑なモデルを扱ってみよう. 次は 2 次元のパラメータの場合である.

ここでも, ある生物種の個体群密度の調査を考える. 生息域内において, ある一定の面積のサイトを K 箇所ランダムに選ぶところは先と同じであるが, 今度は各サイトを T 日間調査するものとする. ここでは, 発見ミスがあることを考慮し, 一頭あたりでの発見率を p とする. ただし, 毎回の観測において個体の数を観察する必要はなく, そのサイト内に個体が生息しているか否かのみを観測する. 調査期間中に各サイト内の個体数に変化は

無いものと仮定する。

先ほどと同様に、それぞれのサイト内に生息する個体数を独立同一なポアソン分布に従う確率変数とし、

$$X_1, X_2, \dots, X_K \stackrel{iid}{\sim} Po(\lambda) \quad (1.52)$$

とおく。ただし、この確率変数は直接観察できない潜在変数である。そこで、サイト k ($k = 1, 2, \dots, K$) における t ($t = 1, 2, \dots, T$) 日目の観察結果を Y_{kt} とする。サイト k における実際の生息数 $X_k = x$ を与えたとき、 $Y_{k1}, Y_{k2}, \dots, Y_{kT}$ は独立同一なベルヌーイ分布 $Bin(1, 1 - q^x)$, $q = 1 - p$ に従う。したがって、サイト k において発見のあった日数 $Y_k = \sum_{i=1}^T Y_{ki}$ の条件付確率分布は

$$Y_k |_{X_k=x} \sim Bin(T, 1 - q^x) \quad (1.53)$$

となり、観測値 Y_k の確率分布は

$$\begin{aligned} Pr(Y_k = y) &= \sum_{i=0}^{\infty} Pr(Y_k = y | X_k = x) Pr(X_k = x) \\ &= \sum_{i=0}^{\infty} \binom{T}{y} (1 - q^x)^y (q^x)^{T-y} e^{-\lambda} \frac{\lambda^x}{x!} \end{aligned} \quad (1.54)$$

という複合ポアソン分布とよばれる形になる。したがって、 K 個のサイトからの尤度はこの積となる。この場合、無限級数を綺麗にまとめることができないため、近似的な計算をすることになる（ただし、次章のベイズ法を用いると、この複合ポアソン分布を厳密に扱うことができる）。

ここでは、 $\lambda = 3$, $p = 0.2$ と仮定して、 $K = 50$ 個のサイトを 10 回ずつ訪れる場合を考える。この想定の下でシミュレーションデータを生成し、その観測値を基に対数尤度の最適化を行う。

```

1 # Parameter setting
2 set.seed(2018)
3 lambda <- 3
4 p <- 0.2; q<-1-p
5 K <- 50
6 T <- 10
7
8 # Data generation
9 x <- rpois(K, lambda)
10 y <- numeric(K)
11 for(k in 1:K) y[k] <- rbinom(1, T, (1-q)^x[k])
12 y

```

(occupancy.data.r)

```
> y
[1] 0 0 1 3 0 2 0 2 0 0 0 0 0 0 0 0 1 0 0 0 0 0 4 0 0 0
[27] 0 0 0 0 1 0 10 0 0 0 10 4 0 0 0 0 0 0 10 0 1 3 0
```

(occupancy.data.r の実行結果)

幸いパラメータが 2 変数なので、(負の)対数尤度の等高線も描くことができ、その概形をとらえることができる。パラメータの最適値（ここでは負の対数尤度の最小化）を数値的に行うために、しばしばパラメータの変換を行う。ここでは $\lambda > 0$, $0 < p < 1$ であることを考え、それぞれ $\log \lambda$, $\text{logit } p = \log(p/(1-p))$ へと変換し、いずれも $-\infty$ から ∞ の値を自由に動けるようにしている。こうすることで、2つの等高線を比べても分かるように最適値の探索もしやすくなる。

```
1 Len <- 100
2 LL <- array(0, c(Len,Len))
3 lamvec <- seq(0,10,length.out=Len)
4 pvec <- seq(0,1,length.out=Len)
5
6 for(i in 1:Len){
7   for(j in 1:Len){
8     LL[i,j] <- NLL(c(log(lamvec[i]), logit(pvec[j]))) )
9   }
10 }
11 contour(lamvec,pvec,LL,nlevels=50, xlab="lambda", ylab="p")
```

(occupancy.contour1.r)

```
1 Len <- 100
2 LL <- array(0, c(Len,Len))
3 loglamvec <- seq(-3,3,length.out=Len)
4 logitpvec <- seq(-4,3,length.out=Len)
5 for(i in 1:Len){
6   for(j in 1:Len){
7     LL[i,j] <- NLL(c(loglamvec[i], logitpvec[j]), Kmax=20)
8   }
9 }
10 contour(loglvec,logitpvec,LL,nlevels=50, xlab="log lambda", ylab="logit p")
```

(occupancy.contour2.r)

最後に負の対数尤度関数を最適化させる。ここでも `optim` という関数を用いた。最適化後にヘッシアンを取り出し、最尤推定値における観測された Fisher 情報行列とし、その逆行列がパラメータ $\log \lambda$, $\text{logit } p = \log(p/(1-p))$ の推定値の分散共分散行列となる。そこからさらにデルタ法によって基のパラメータ λ, p の標準誤差が計算される。

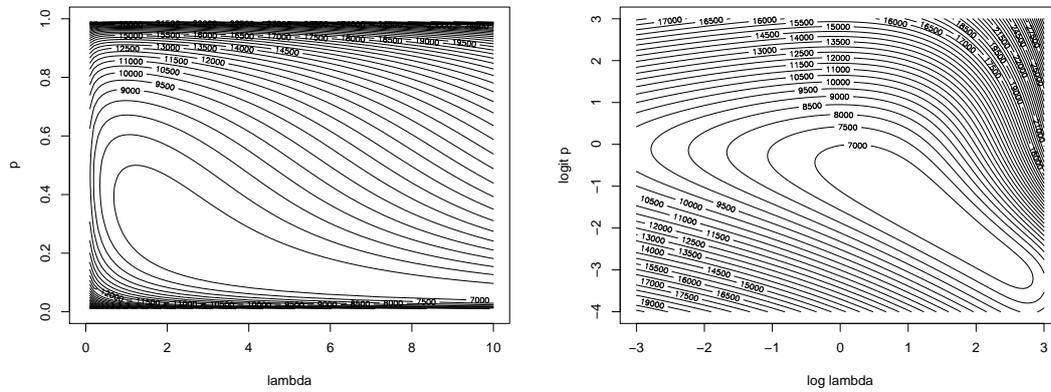


図 1.3: 負の対数尤度関数の等高線図. 左図: (λ, p) の関数として. 右図: $(\log \lambda, \text{logit } p = \log(p/(1-p)))$ の関数として.

```

1  init <- c(0,0)
2  Res <- optim(init, NLL, method="BFGS", hessian=T, Kmax=20)
3  Res
4  points(Res$par[1],Res$par[2], col="red", pch=4)
5
6  lambda.est <- exp(Res$par[1])
7  p.est <- 1/(1+exp(-Res$par[2]))
8  Avar <- solve(Res$hessian)
9  lambda.se <- lambda.est*sqrt(Avar[1,1])
10 p.se <- p.est*(1-p.est)*sqrt(Avar[2,2])
11 data.frame(lambda.est, lambda.se, p.est, p.se)

```

(occupancy.est.r)

```

> data.frame(lambda.est, lambda.se, p.est, p.se)
  lambda.est lambda.se  p.est  p.se
1  2.961073 0.1653984 0.1999782 0.01016644

```

(occupancy.est.r の実行結果)

[一口メモ] 最尤法では、観測値の数が多くなると、推定値が真値に近づき、かつ漸近分散を通して誤差評価ができるなどのメリットがあった。また推定、誤差評価、信頼区間の構成、仮説検定、モデル選択と一貫した推測のフレームワークも提供してくれる。しかしながら、良いことばかりではなく、推定するパラメータが多い場合、局外パラメータと呼ばれるパラメータが標本数の増加とともに増える場合（例えば、Neyman-Scott 問題、errors-in-variables model）等では、通常の尤度を修正をしなければ無視できない偏りが生じたり、推定精度が他の推定方法と比較して劣る場合もある。必要に応じて、後の章でこれらの点についても実例を通してみていく。

1.3.3 パラメータの信頼区間と仮説検定

信頼域 信頼域の構成は、仮説検定の裏返しによって求められる。この時、最尤原理に大きく依存した検定統計量に尤度比検定統計量がある。これは

$$T_1 = -2 \log \frac{L(\theta_0)}{L(\hat{\theta})} = 2(l(\hat{\theta}) - l(\theta_0))$$

すなわち、任意の θ に対して、

$$P_{\theta}(T_i \leq \chi_{d,\alpha}^2) = 1 - \alpha \quad (i = 1, 2, 3)$$

であるから

$$C_i(Y) = \{\theta \in \Theta; T_i \leq \chi_{d,\alpha}^2\} \quad (i = 1, 2, 3)$$

が信頼域となる。なお、仮説検定と同様に、水準は厳密には維持されていない可能性がある。この求め方では、 d が大きい時には全く現実的ではない。但し、 T_3 については工夫により比較的容易に信頼域を求める事ができる。例えば、観測された Fisher 情報量行列 (observed Fisher information matrix) を

$$J(\theta) = -\frac{\partial^2}{\partial \theta \partial \theta^t} l(\theta)$$

で定義する。この時、任意の θ に対して、

$$I(\hat{\theta}), J(\theta), J(\hat{\theta}) \longrightarrow I(\theta) \quad \text{in prob}$$

が成り立つ。従って、 T_3 における $I(\theta)$ をそれに確率収束する統計量 $I(\hat{\theta}), J(\theta), J(\hat{\theta})$ で置き換えても漸近分布は変わらないから、例えば、

$$T_3 = (\hat{\theta} - \theta)^t I(\hat{\theta})(\hat{\theta} - \theta) \quad \text{or} \quad T_3 = (\hat{\theta} - \theta)^t J(\hat{\theta})(\hat{\theta} - \theta)$$

を用いれば、信頼域の計算は非常に容易となる。

仮説検定 いま、帰無仮説 $H_0 : \theta = \theta_0$ に対して、対立仮説 $H_1 : \theta \neq \theta_0$ を考える。この時、最尤原理に大きく依存した検定統計量に尤度比検定統計量がある。これは

$$T_1 = -2 \log \frac{L(\theta_0)}{L(\hat{\theta})} = 2(l(\hat{\theta}) - l(\theta_0))$$

で定義される。最尤原理では最も支持される $\hat{\theta}$ から θ_0 がどれくらい離れているかを測っている事になる (K-L との対応も入れておく)。 n が十分大きい時、 T_1 は漸近的に χ^2 分布に従うことから、棄却域は

$$R_1(Y) = \{Y; T_1 \geq \chi_{d,\alpha}^2\}$$

で与えられる。但し、 $\chi_{d,\alpha}^2$ は自由度 d の χ^2 分布における $100(1 - \alpha)\%$ 点である。

1.3.4 プロファイル尤度

いま、興味あるパラメータ ϕ を固定したときの残りのパラメータ λ の最尤推定量が明示的に $\hat{\lambda}(\phi)$ として求まるとする。 $\hat{\lambda}(\phi)$ を改めて元の尤度 $L(\phi, \lambda)$ に代入することによって得られる ϕ の関数

$$L_p(\phi) = L(\phi, \hat{\lambda}(\phi)) \quad (1.55)$$

を ϕ のプロファイル尤度という。同様にプロファイル対数尤度も $l_p(\phi) = \log L_p(\phi)$ で定義される。 $\hat{\phi}$ がパラメータ空間の内点である時、 $L(\phi, \lambda)$ による最尤推定量と $L_p(\phi)$ の最大化による推定量は一致する。従って、通常は推定値自体には変化はない。ここでは、パラメータ ϕ の信頼域の構成についてのみ述べることにする。

信頼域の構成および検定は、形式的に通常の場合と同様に尤度比統計量を用いて行うことができる。いま、プロファイル尤度の比が

$$2(l_p(\hat{\phi}) - l_p(\phi)) = 2(l(\hat{\phi}, \hat{\lambda}) - l(\phi, \hat{\lambda}(\phi)))$$

と表せる。従って、

$$2(l_p(\hat{\phi}) - l_p(\phi)) \longrightarrow \chi_{d_\phi}^2 \quad \text{in dist}$$

が成り立ち、

$$C_1(Y) = \{\phi; 2(l_p(\hat{\phi}) - l_p(\phi)) \leq \chi_{d_\phi}^2\}$$

により、 ϕ の信頼域を構成できる。

例題：混合率？

疲れた (Aug 12)

1.4 モデル選択

(to come)

索引

確率収束, 8
観測された Fisher 情報行列, 10

最小 2 乗法, 7
最尤推定値, 3
最尤推定量, 3
最尤法, 2

スコア関数, 4

線形回帰モデル, 7

対数尤度関数, 3

デルタ法, 13

標準誤差, 11

分布収束, 9

尤度関数, 3