

私のための「頻度論」統計学入門

阿部 興*

2020 年 2 月 7 日

記法

$\log(x)$ は自然対数を表す.

1 確率の話

ここでは主に大数の法則, 中心極限定理, カルバック・ライブラ情報量について勉強する.

1.1 大数の法則

いろいろな本に書いてある.

1.2 中心極限定理

中心極限定理の証明には, 特性関数やモーメント母関数を使うことが多い.

テイラー展開を使ったより簡単な証明が, 黒木 (2017)[1] にある.

*「あべ こう」と読む

1.3 カルバック・ライブラ情報量

統計的推測のために重要な役割を果たす量の 1 つにカルバック・ライブラ情報量がある. カルバック・ライブラ情報量 $D_{KL}(p||q)$ は離散型の確率関数 $p(x)$ と $q(x)$ に対しては,

$$D_{KL}(p||q) = - \sum_x p(x) \log \frac{q(x)}{p(x)} \quad (1)$$

連続型の確率密度関数 $p(x)$ と $q(x)$ に対しては,

$$D_{KL}(p||q) = - \int p(x) \log \frac{q(x)}{p(x)} dx \quad (2)$$

と定義される. 赤池 (1980) [2] に習って, カルバック・ライブラ情報量の直感的な意味を考える. 結論を先取りすると, カルバック・ライブラ情報量は「サンプルを生成した分布が $q(x)$ のとき, 経験分布がほぼ $p(x)$ となる確率の対数のサンプルサイズ分の 1 の符号反転」と解釈できる. このカギカッコの中身の意味がわかるようになることがこの節の目標である.

状態 $i = 1, \dots, M$ がそれぞれ q_i の確率で生起する分布と, 状態 $i = 1, \dots, M$ がそれぞれ p_i の確率で生起する分布を考える. この分布からのサンプルを N 個観測して, 状態 $i = 1, \dots, M$ が生起した回数をそれぞれ N_1, \dots, N_M とする. $N = \sum_{i=1}^M N_i$ である. N_1, \dots, N_M のような観測が得られる確率は,

$$W = \frac{N!}{N_1! \dots N_M!} q_1^{N_1} \dots q_M^{N_M} \quad (3)$$

と表せる. (このような分布は多項分布と呼ばれる.)

ここでスターリングの公式

$$\log N! \approx N \log N - N \quad (4)$$

を使って $\log W$ を近似すると

$$\log W \approx (N \log N) - \sum_{i=1}^M (N_i \log N_i - N_i) + \sum_i 1^M \log q_i \quad (5)$$

$$= N \log N - \sum_{i=1}^M N_i (\log N_i - q_i) \quad (6)$$

$$= - \sum_{i=1}^M N_i (\log N_i - q_i - \log N) \quad (7)$$

$$= - \sum_{i=1}^M N_i (\log \frac{N_i}{N} - q_i) \quad (8)$$

$$= -N \sum_{i=1}^M \frac{N_i}{N} (\log \frac{N_i}{N} - q_i) \quad (9)$$

と整理できる. $p_i = N_i/N$ とおくと

$$\log W \approx -N \sum_{i=1}^M p_i \log \frac{p_i}{q_i} \quad (10)$$

という結果を得る.

あらためて考えると $p_i = N_i/N$ は経験的に推定された確率と解釈できる. q_i は真の確率であったことを思い出すと、 W はサンプルを生成した真の分布が q_i のときに p_i のように振る舞う確率と解釈できる. $\log W$ を N で割り、符号を反転させると、 $-\sum_{i=1}^M p_i \log \frac{q_i}{p_i}$ となる. この量は「母集団分布が q_i のとき経験分布がほぼ p_i となる確率の対数のサンプルサイズ分の 1 の符号反転」と解釈できる.

カルバック・ライブラ情報量は次の性質を満たすため、分布の近さを測る指標となる.

$$D_{KL}(p||q) \geq 0, \quad (11)$$

かつ、 $D_{KL}(p||q) = 0$ となるのは $p(x) = q(x)$ のときに限られる.

注意. 定義をよく見ればわかるように、カルバック・ライブラ情報量では $D_{KL}(p||q) = D_{KL}(q||p)$ は成り立たない.

2 最尤法

本節では最尤法と呼ばれる方法の性質について述べる。これについて理解するために、フィッシャー情報量と呼ばれる量が重要になるため、先にフィッシャー情報量についての性質を述べる。

2.1 パラメータが1つの場合

2.1.1 スコア関数とフィッシャー情報量

パラメータ θ を持つ確率（密度）関数 $p(x|\theta_0)$ について、スコア関数 $S(\theta)$ を次のように定義する。

$$S(\theta) = \frac{d}{d\theta} \log p(x|\theta). \quad (12)$$

スコア関数の $p(x|\theta)$ による平均は 0 である。

$$\int_{-\infty}^{\infty} \frac{d}{d\theta} \log p(x|\theta) p(x|\theta) dx \quad (13)$$

$$= \int_{-\infty}^{\infty} \frac{\frac{d}{d\theta} p(x|\theta)}{p(x|\theta)} p(x|\theta) dx \quad (14)$$

$$= \int_{-\infty}^{\infty} \frac{d}{d\theta} p(x|\theta) dx \quad (15)$$

$$= \frac{d}{d\theta} \int_{-\infty}^{\infty} p(x|\theta) dx \quad (16)$$

$$= 0. \quad (17)$$

従い、スコア関数の分散はスコア関数の 2 乗の平均に等しい。

フィッシャー情報量を

$$I(\theta) = - \int_{-\infty}^{\infty} \left(\frac{d^2}{d\theta^2} \log p(x|\theta) \right) p(x|\theta) dx \quad (18)$$

と定義すると,

$$I(\theta) = - \int_{-\infty}^{\infty} \frac{d}{d\theta} \left(\left(\frac{\frac{d}{d\theta} p(x_i|\theta)}{p(x|\theta)} \right) p(x|\theta) \right) dx \quad (19)$$

$$= - \int_{-\infty}^{\infty} \left(\left(\frac{\frac{d^2}{d\theta^2} p(x_i|\theta)}{p(x|\theta)} - \frac{(\frac{d}{d\theta} p(x_i|\theta))^2}{p(x|\theta)^2} \right) p(x|\theta) \right) dx \quad (20)$$

$$= - \int_{-\infty}^{\infty} \left(\frac{\frac{d^2}{d\theta^2} p(x_i|\theta)}{p(x|\theta)} \right) p(x|\theta) dx + \int_{-\infty}^{\infty} \left(\frac{(\frac{d}{d\theta} p(x_i|\theta))^2}{p(x|\theta)^2} \right) p(x|\theta) dx \quad (21)$$

スコア関数のときと同様, 第1項は消える. 第2項は

$$\int_{-\infty}^{\infty} \left(\frac{d}{d\theta} \log p(x|\theta) \right)^2 p(x|\theta) dx \quad (22)$$

と等しい. これはスコア関数の2乗の平均になっている. すなわち, スコア関数の分散はフィッシャー情報行列と等しいことがわかった.

2.1.2 最尤推定量の性質

サンプル x_i ($i = 1, \dots, n$) が, 独立に同一の確率 (密度) 関数 $p(x|\theta_0)$ を持つ分布から得られたとする¹. ここで θ は確率 (密度) 関数のパラメータである². このようなデータに対し, 統計モデル $p(x|\theta)$ を考え, 未知パラメータの θ を推定したい.

まず, 次の対数尤度比関数 $l_n(\theta)$ を考える.

$$l_n(\theta) = \log \left(\frac{\prod_{i=1}^n p(x_i|\theta)}{\prod_{i=1}^n p(x_i|\theta_0)} \right) = \sum_{i=1}^n \log \left(\frac{p(x_i|\theta)}{p(x_i|\theta_0)} \right). \quad (23)$$

これをサンプルサイズ (標本の大きさ) n で割ると大数の法則により,

$$\lim_{n \rightarrow \infty} l_n(\theta)/n = \int_{-\infty}^{\infty} p(x|\theta_0) \log \left(\frac{p(x|\theta)}{p(x|\theta_0)} \right) dx \quad (24)$$

¹ 「独立に同一の確率分布に従う」ことを i.i.d. (independent and identically distributed) と略すことがある.

² このパラメータのことを母数と呼ぶことがあるが, 母数という語は誤解を招くことが多いので, 本稿ではあまりつかわない.

となる. 右辺はサンプルを生成した分布と, 統計モデルのカルバック・ライブラ情報量の -1 倍となっている. そのため, サンプルを生成した分布と統計モデルのカルバック・ライブラ情報量を最小にするためには $l_n(\theta)$ を最大にすればよいことが予想される. $l_n(\theta)$ の式を少し変形する.

$$l_n(\theta) = \sum_{i=1}^n \log p(x|\theta) dx - \sum_{i=1}^n \log p(x_i|\theta_0) dx. \quad (25)$$

右辺第2項は, サンプルを生成した分布のみによって定まる量であり, 統計モデルのパラメータ θ の選び方に依存しない. よって, $l_n(\theta)$ を最大にするためには, 第1項を最大にする θ を探せばよい. これが最尤法と呼ばれる推定方法のアイデアである.

$l_n(\theta)$ を最大化する θ は, 確率変数としてどのような振る舞いをするだろうか. そのことを調べるために, テイラー展開を使う. 関数 $f(x)$ の x_0 のまわりでのテイラー展開は,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \dots \quad (26)$$

であった. テイラー展開についてよく知らない場合は (なにかいい本) を参考にとするとよい. $l_n(\theta)$ の θ_0 の周りでのテイラー展開は,

$$l_n(\theta) = l_n(\theta_0) + l'_n(\theta_0)(\theta - \theta_0) + \frac{1}{2}l''_n(\theta_0)(\theta - \theta_0)^2 + \dots \quad (27)$$

である. $h/\sqrt{n} = \theta - \theta_0$ と置くと,

$$l_n(\theta_0 + h/\sqrt{n}) = l_n(\theta_0) + \frac{l'_n(\theta_0)}{\sqrt{n}}h + \frac{1}{2n}l''_n(\theta_0)h^2 + \dots \quad (28)$$

となる. h を \sqrt{n} で割ったのは後の計算の便宜のためである. $l_n(\theta_0) = \log 1 = 0$ であるから,

$$l_n(\theta_0 + h/\sqrt{n}) = \frac{l'_n(\theta_0)}{\sqrt{n}}h + \frac{1}{2n}l''_n(\theta_0)h^2 + O(1/\sqrt{n}) \quad (29)$$

と書ける. ここで O はランダウの記号である. ランダウの記号については (なにかいい本) を参考にとするとよい. $l_n(\theta)$ の定義に戻ると,

$$\begin{aligned} l_n(\theta_0 + h/\sqrt{n}) &\approx \frac{1}{\sqrt{n}} \left[\sum_{i=1}^n \left(\frac{d}{d\theta} \log p(x_i|\theta)|_{\theta=\theta_0} \right) \right] h + \frac{1}{2n} \left[\sum_{i=1}^n \left(\frac{d^2}{d\theta^2} \log p(x_i|\theta)|_{\theta=\theta_0} \right) \right] h^2. \end{aligned} \quad (30)$$

スコア関数とフィッシャー情報行列についての性質を思い出すと、大数の法則と中心極限定理より n が十分大きいとき、標準正規分布に従う確率変数 Z を用いて、

$$l_n(\theta_0 + h/\sqrt{n}) \approx Z\sqrt{I(\theta_0)}h - \frac{I(\theta_0)h^2}{2} \quad (31)$$

$$= -\frac{I(\theta_0)}{2} \left(h - \frac{Z}{\sqrt{I(\theta_0)}} \right)^2 + \frac{Z^2}{2} \quad (32)$$

という近似が成り立つ。 $\theta = \theta_0 + h/\sqrt{n}$ であったので、

$$l_n(\theta) = -\frac{I(\theta_0)}{2} \left(\sqrt{n}(\theta - \theta_0) - \frac{Z}{\sqrt{I(\theta_0)}} \right)^2 + \frac{Z^2}{2} \quad (33)$$

右辺は2次関数であり、 $\sqrt{n}(\theta - \theta_0) = Z/\sqrt{I(\theta_0)}$ のとき最大になる。よって $l_n(\theta)$ を最大にするよう θ を決めると、 $\sqrt{n}(\theta - \theta_0)$ は平均 0、分散 $I(\theta_0)^{-1}$ の正規分布に従う。これが最尤法の基礎である。最尤法により推定された θ を最尤推定量と呼ぶ。最尤推定量は確率変数であるから、サンプルを生成した分布のパラメータそのものではない。そこで最尤推定量は $\hat{\theta}$ などの記号を用いて、パラメータと区別する。

より直感的に述べると、最尤推定量 $\hat{\theta}$ はサンプルサイズが十分大きいとき、平均 θ_0 、分散 $I(\theta_0)^{-1}/n$ の正規分布に従うということである。

最尤推定の手順.

1. データ X を生成した分布に対し、同時確率（密度）関数 $p(X|\theta)$ を適当に与える.
2. $p(X|\theta)$ を θ の関数とみて、 $p(X|\theta)$ を最大化する θ をなんらかの方法で探し、推定量とする.

$p(X|\theta)$ を θ の関数とみるときは尤度関数と呼ばれる。その対数をとったもの $\log p(X|\theta)$ を対数尤度関数と呼ぶ。対数を取ることは多くの場合、計算をととても便利にする。

2.1.3 例題

まずは二項分布、ポアソン分布、指数分布などで練習すると良い。

2.2 パラメータが複数の場合

対象とする確率分布がパラメータを複数持つ場合, 多変数のテイラー展開を用いることで, パラメータが1つの場合と同様の議論を展開することができる.

2.3 間違ったモデルで最尤推定すること

上記ではサンプルを生成した分布と統計モデルが, パラメータのとり方によっては厳密に一致する場合を論じた. しかし, 現実にはサンプルを生成した分布は未知であり, 統計モデルは分析者が設定する. そのため, 統計モデルによって, サンプルを生成した分布が実現可能かどうかはわからない.

最尤推定は真の分布とモデルの間のカルバック・ライブラ情報量を経験的に最小化する方法であったことを思い出すと, 最尤法は「間違ったモデル」を選んでも, 選んだ範囲内で (カルバック・ライブラ情報量の意味で) 適切な推定を行うことが予想される.

しかし, このことを一般的に保証するような定理は今のところないと思う.

2.3.1 例題

サンプルを生成した分布が確率密度関数

$$g(x) = \frac{1}{6}x^3 \exp(-x) \quad (34)$$

を持つとする. (この分布はガンマ分布と呼ばれる分布の一例である.) 統計モデルを正規分布とする.

サンプルを生成した分布と統計モデルの間のカルバック・ライブラ情報量は,

$$E_g(-\log 6 + 3 \log X - X) - \frac{1}{2}E_g\{\log(2\pi\sigma^2)\} + \frac{(X - \mu)^2}{\sigma^2} \quad (35)$$

である. ここで, $E_g[X] = \int xg(x) dx$.

(途中計算は省略して)

$\hat{\mu} = 4$, $\hat{\sigma}^2 = 4$ で, 真の分布とモデルの間のカルバック・ライブラ距離は最小となる. R を使って最尤推定のシミュレーションを試みる.


```

MLEnorm_sim <- function(i,n){
  x <- rgamma(n,4,1)
  muhat <- mean(x)
  n <- length(x)
  s2hat <- n*var(x)/(n-1)
  c(muhat,s2hat)
}

library(parallel)
res10 <- t(simplify2array(mclapply(1:10000,MLEnorm_sim,n=10,
  mc.cores = detectCores()))))
res100 <- t(simplify2array(mclapply(1:10000,MLEnorm_sim,n=100,
  mc.cores = detectCores()))))
res1000 <- t(simplify2array(mclapply(1:10000,MLEnorm_sim,n=1000,
  mc.cores = detectCores()))))

boxplot(cbind("n=10"=res10[,1],"n=100"=res100[,1],
  "n=1000"=res1000[,1]),main=expression(hat(mu)))
abline(h=4,lty=2)

boxplot(cbind("n=10"=res10[,2],"n=100"=res100[,2],
  "n=1000"=res1000[,2]),main=expression(hat(sigma^2)))
abline(h=4,lty=2)

```

3 カイ二乗分布を使った検定

Z を標準正規分布に従う確率変数とすると Z^2 は自由度 1 のカイ二乗分布に従う. 自由度 r のカイ二乗分布とは, 以下の確率密度関数を持つ分布である.

$$f(x) = \frac{1}{2^{r/2}\Gamma(r/2)} x^{r/2-1} \exp(-x/2). \quad (36)$$

また Z_1, Z_2, \dots, Z_r が独立でそれぞれ標準正規分布に従うとき, $Z_1^2 + Z_2^2 + \dots + Z_r^2$ は自由度 r のカイ二乗分布に従う.

(33) 式の結果を思い出すと、対数尤度比関数 $l_n(\theta)$ が最大になるとき、その最大値の 2 倍は Z^2 であった。このことから、尤度関数を用いた検定にカイ二乗分布が使えることがわかる。

3.1 Wilks の定理

あとで書く。

3.2 2×2 分割表の独立性の検定

様々な分野で表 1 のようなカウントデータの表を考えることがある。このとき、各行（暴露の有無）と各列（疾病の有無）は独立か、という問いに関心があることが多い。ここでいう独立とは、...

表 1: 2×2 分割表.

	疾病あり	疾病なし	合計
暴露あり	a_{11}	a_{12}	N_1
暴露なし	a_{21}	a_{22}	N_2
合計	M_1	M_2	N

このようにシンプルな例でも、統計モデルは複数考えることができる。以下では、この分割表を $A = (a_{ij})$ と表記する。

3.3 二項分布モデル

a_{i1} がそれぞれ独立にパラメータ q_i の二項分布に従うとする。 $Q = (q_i)$ とおく。 A が生じる確率 $p(A|Q)$ は、

$$p(A|Q) = \prod_{i=1}^2 \frac{N_i!}{a_{i1}!a_{i2}!} (q_i^{a_{i1}}(1 - q_i)^{a_{i2}}) \quad (37)$$

とかける。この場合、 q_i の最尤推定量は、 $\hat{q}_i = a_{i1}/N_i$ である。

分割表の各行と各列は独立であるとするモデルを、パラメータが $q_0 = q_1 = q_2$ という条件を満たしているモデルだとする。この場合、 q_0 の最尤推定量は、 $\hat{q}_0 = M_1/N$ である。

次にこの2つのモデルの対数尤度比統計量 G_0 を考える. 3.1 節で述べた通り, G_0 はサンプルサイズが十分大きいとき, 自由度1のカイ二乗分布に近似的に従う.

$$G_0 = 2 \left(\sum_{i=1}^2 \{a_{i1} \log(\hat{q}_i) + a_{i2} \log(1 - \hat{q}_i)\} - \right. \quad (38)$$

$$\left. \sum_{i=1}^2 \{a_{i1} \log(\hat{q}_0) + a_{i2} \log(1 - \hat{q}_0)\} \right) \quad (39)$$

$$= 2 \left(\sum_{i=1}^2 \left\{ a_{i1} \log \frac{\hat{q}_i}{\hat{q}_0} + a_{i2} \log \frac{1 - \hat{q}_i}{1 - \hat{q}_0} \right\} \right) \quad (40)$$

ここで $\hat{q}_i/\hat{q}_0 = a_{i1}/(N_i M_i/N)$, $(1 - \hat{q}_i)/(1 - \hat{q}_0) = a_{i2}/(N_i M_i/N)$ と書ける. この統計量 G_0 を用いた検定を G 検定と呼ぶことがある. また a_{ij} を観測度数, $N_i M_i/N$ や $N_i M_i/N$ を理論度数と呼ぶことがある. 2×2 分割表の独立性の検定において, 理論度数を表2のように定めると, 統計量 G は次のようにも書ける.

$$G = 2 \left(\sum_{i,j} a_{ij} \log \frac{a_{ij}}{E_{ij}} \right). \quad (41)$$

表 2: 2×2 分割表の独立性の検定における理論度数.

	疾病あり	疾病なし	合計
暴露あり	$E_{11} = N_1 M_1/N$	$E_{12} = N_1 M_2/N$	N_1
暴露なし	$E_{21} = M_1 N_2/N$	$E_{22} = M_2 N_2/N$	N_2
合計	M_1	M_2	N

ところで, $\log(1+x)$ を0のまわりでテイラー展開することで次の式が成り立つ.

$$(1+x) \log(1+x) = (1+x) \left(x - \frac{x^2}{2} + O(x^3) \right) \quad (42)$$

$$= x + \frac{x^2}{2} + O(x^3) \quad (43)$$

2つめの等号において, かっこを展開すると x^3 についての項が出てくるが, $O(x^3)$ に吸収される.

この式を用いると, 統計量 G の式は, 以下のように近似できる.

$$2 \left(\sum_{i,j} a_{ij} \log \frac{a_{ij}}{E_{ij}} \right) \quad (44)$$

$$= 2 \left(\sum_{i,j} E_{ij} \left(1 + \frac{a_{ij} - E_{ij}}{E_{ij}} \right) \log \left(1 + \frac{a_{ij} - E_{ij}}{E_{ij}} \right) \right) \quad (45)$$

$$\approx 2 \left(\sum_{i,j} E_{ij} \left(\frac{a_{ij} - E_{ij}}{E_{ij}} - \frac{(a_{ij} - E_{ij})^2}{E_{ij}^2} \right) \right) \quad (46)$$

$$= 2 \left(\sum_{i,j} \frac{(a_{ij} - E_{ij})^2}{E_{ij}} \right). \quad (47)$$

最後の等号においては, $\sum_{i,j} E_{ij} = \sum_{i,j} a_{ij} = N$ を用いた. この統計量を χ_0^2 とおく. χ_0^2 は G_0 とサンプルサイズが十分大きいとき等しい. すなわち, χ_0^2 はサンプルサイズが自由度 1 のカイ二乗分布に近似的に従う.

単にカイ二乗検定というときは, この χ_0^2 統計量を用いた検定を指すことが多い.

ではカイ二乗検定の手順を具体的に示そう.

カイ二乗検定の手順. データから求めた検定統計量 (この場合 G_0 や χ_0^2) 以上の値が, 自由度 1 のカイ二乗分布から出る確率を調べる. その調べた確率を p 値と呼ぶ. p 値が小さすぎるとき, 分割表の各行と各列は独立でないと判断する.

どの程度の確率であれば, 小さすぎると判断するかは事前に定めておく. この事前に定めた閾を有意水準と呼ぶ.

3.3.1 ポアソン分布モデル

注意点. 統計的仮説検定は確率版背理法に例えられることもあるが, 現実のデータを扱う以上, 背理法のようにすっきりとはいかない. 具体的にどのような難しさがあるかは 節で改めて詳しく述べる.

3.3.2 カイ二乗検定のシミュレーション

χ_0^2 は G_0 とサンプルサイズが十分大きいとき等しいと述べたが, サンプルサイズが小さいときには大きく性質がことなる. どの程度のサンプル

サイズるとき、どのような性質を持つかを調べるには、コンピューターを使ってシミュレーションするのが早い。

よい仮説検定の第1条件は、まずなによりも α エラーをコントロールできること、第2の条件は検出力が高いことである。

G 検定はサンプルサイズの小さいとき、実際の α エラーが名目上の有意水準を上回ることが多い。カイ二乗検定は概ね正確な p 値を与えることが多い。

注意点. 本説の議論は、一般の $r \times c$ 分割表にも拡張できる。しかし、 $r \times c$ 分割表をカイ二乗検定で分析するのは、私があまり好きではないので、本稿では扱わない。 r と c が 2 より大きいとき、 $r \times c$ 分割表の独立性のカイ二乗検定では、行と列が独立でないことは言えても、どの行とどの列に特に大きな関係があるかはわからない。そこでカイ二乗検定を行った後に、残差分析と呼ばれる分析を行うことがある。しかし、このように検定を何段階も行うことの、多重比較のような問題を考え出すと話がややこしくなるように思う。連続量を大、中、小のように分けて表にした場合は、元の連続量に対する統計モデルを考えたほうがよいように思う。また、順序尺度の量を集計した場合は、順序尺度に対する統計モデルを考えたほうがよいように思う。順序尺度に対する統計モデルは、7 節で扱う予定である。

4 ワルド検定とワルド信頼区間

まず、次の定理を紹介する。

定理 1. $\hat{\theta}$ が最尤推定量であるとき、関数 $g(\theta)$ の最尤推定量は $g(\hat{\theta})$ である。これを最尤推定量の不変性という。

ただし、 $g(\theta)$ が 1 対 1 の変換になっていないとき、この定理は成り立たない。

自明な事実を述べたにすぎないと思われると思うので、この定理はなにがうれしいか補足する。

例えば正の値しか取らないパラメータ σ を最優推定したいとき、制約付きの最適化をするのは面倒である。このようなとき、 $\sigma = \exp(\rho)$ において、 ρ を最尤推定し、 $\hat{\sigma} = \exp(\hat{\rho})$ として構わない。

4.1 ワルド検定

2.1.2 節で述べた通り、サンプルサイズが十分大きいとき、最尤推定量は平均 θ 、分散 $I(\theta)^{-1}/n$ の正規分布で近似できた。ここで $I(\theta)$ はフィッシャー情報量である。しかし、仮に統計モデルとしてデータを生成した分布を実現可能な分布を選んだとしても³、正しいパラメータ θ は不明であり、解析的な期待値の計算も難しいことが多い。

そこでサンプル x_i ($i = 1, \dots, n$) が得られたときの、統計モデル $p(x|\theta)$ の観測情報量 $I_o(\theta)$ を次のように定義する。

$$I_o(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log p(x_i|\theta). \quad (48)$$

サンプルサイズ n が十分大きいとき、 $I_o(\theta)$ は $I(\theta)$ に近づく。これにより、最尤推定量は分散 $I_o(\hat{\theta})^{-1}/n$ の正規分布で近似できる。

4.2 ワルド信頼区間

ワルド信頼区間を次のように定義する。

5 回帰型の統計モデル

5.1 ロジスティック回帰

5.2 ポアソン回帰

6 モデル選択

6.1 クロスバリデーション

注意点. ディープラーニングのパッケージなどでは、バリデーション損失を簡単にモニタできるようになっていることが多い。これは便利であるが、弊害もあると感じている。例えばバリデーション損失をモニタしながら最適化を行い、そのバリデーション損失をそのまま予測精度として報

³一般には、データを生成した分布を実現可能な分布ということはとても難しい。その方法はわかっていない。

告した場合, 間接的にバリデーション用のデータにフィットするように最適化を行っているのと同じことが起こる気がする. チューニングパラメータの多いモデルの場合, 理想的には, チューニングパラメータを決めるためのバリデーションデータでチューニングを行い, 最終的な予測精度の評価はまた別のバリデーションデータで行うほうが良いと思う. しかしこの方法ではデータが大量に必要なことになる.

7 いくつかの統計モデルのカatalog

統計モデルを紹介した楽しい本に松浦 や 須山 がある. ここではこれらの本に記載されていない統計モデルを紹介する. これにより自分の手で統計モデルを作るということがどのようなことか, その雰囲気の一端を感じてもらいたい.

参考文献

- [1] 黒木玄. 2017-06-04 Lindeberg の Taylor 展開のみを使った中心極限定理の証明.pdf. <https://genkuroki.github.io/documents/mathtodon/>.
- [2] 赤池弘次. (1980). エントロピーとモデルの尤度 (〈講座〉物理学周辺の確率統計). 日本物理学会誌, 35(7), 608-614.