私のための統計学入門

阿部 興 *

2020 年 2 月 6 日

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ので省略する.

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.

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 $D_{\mathrm{KL}}(p\|q)$ は離散型の確率関数 $p(x)$ と $q(x)$ に対しては,

$$D_{\mathrm{KL}}(p\|q) = -\sum_{x} p(x) \log \frac{q(x)}{p(x)} \tag{1}$$

連続型の確率密度関数 $p(x)$ と $q(x)$ に対しては,

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

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

*「あべ こう」と読む

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 $i = 1, \ldots, M$ がそれぞれ q_i の確率で生起する分布と, 状態 $i = 1, \ldots, M$ 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 $i = 1, \ldots, M$ が生起した回数をそれぞれ $N_1, \ldots, N_M$ とする. $N = \sum_{i=1}^{M} N_i$ である. $N_1, \ldots, N_M$ のような観測が得られる確率は,

$$W = \frac{N!}{N_1! \cdots N_M!} q_1^{N_1} \cdots q_M^{N_M} \tag{3}$$

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

$$\log N! \approx N \log N - N \tag{4}$$

を使って $\log W$ を近似すると

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

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

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

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

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

と整理できる. $p_i = N_i/N$ とおくと

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

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.

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

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

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.

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

スコア関数の $p(x|\theta)$ による平均は 0 である.

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

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

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

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

$$= 0. \tag{17}$$

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

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

と定義すると,

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

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

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

スコア関数のときと同様, 第 1 項は消える. 第 2 項は

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

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 $(i = 1, \ldots n)$ 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, 次のような関数 $l_n(\theta)$ を考える.

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

これをサンプルサイズ（標本の大きさ）で割ると大数の法則により,

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

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.

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

[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.

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,

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

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,

$$l_n(\theta) = l_n(\theta_0) + l_n'(\theta_0)(\theta - \theta_0) + \frac{1}{2}l_n''(\theta_0)(\theta - \theta_0)^2 + \cdots \tag{27}$$

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

$$l_n(\theta_0 + h/\sqrt{n}) = l_n(\theta_0) + \frac{l_n'(\theta_0)}{\sqrt{n}}h + \frac{1}{2n}l_n''(\theta_0)h^2 + \cdots \tag{28}$$

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,

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

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,

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

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,

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

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

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

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

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 $I(\theta_0)^{-1}/\sqrt{n}$ 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.

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, 多変数のテー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われる.

2.3.1 例題

サンプルを生成した分布が確率密度関数

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

を持つとする. （この分布はガンマ分布と呼ばれる.）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,

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

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してみる.

```
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 2 乗検定

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.pdfhttps://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.