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　カルバック・ライブラ情報量

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 $tex : p_i$ の確率で生起する分布を考える.

この 2 つの分布間のカルバック・ライブラ情報量は、

$$\sum_{i=1}^{M} p_i \log \frac{p_i}{q_i}$$

と定義されます。

この量は「母集団分布が $tex : q_i$ 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. この一文の意味がわかるようになることがこの節の目標である. さて、状態 $tex : i = 1, \ldots, M$ がそれぞれ q_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 $N_1, \ldots, N_M$ とします。$N = \sum_{i=1}^{M} N_i$ です。$N_1, \ldots, N_M$ のような観測が得られる確率は

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

です。（要は多項分布です。）

ここでスターリングの公式

$$\log N! \approx N \log N - N$$

*「あべ こう」と読む

を使って $tex : \log W$ を近似すると

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

$$= N \log N - \sum_{i=1}^{M} N_i(\log N_i - q_i) \tag{2}$$

$$= -\sum_{i=1}^{M} N_i(\log N_i - q_i - \log N) \tag{3}$$

$$= -\sum_{i=1}^{M} N_i(\log \frac{N_i}{N} - q_i) \tag{4}$$

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

と整理できる。$p_i = N_i/N$ とおくと

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

という結果を得る.

あらためて考えると $tex : p_i = N_i/N$ は経験的に推定された確率と解釈できます。$tex : q_i$ は真の確率であったことを思い出すと、W は真の分布が q_i のときに p_i のように振る舞う確率と解釈できます。$\log W]$ を N で割って、符号を反転させると、$\sum_{i=1}^{M} p_i \log \frac{p_i}{q_i}$ となります。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」と解釈できます。

以上の考察からカルバック・ライブラ情報量を次のように定義する.

$$D(p\|q) \tag{6}$$

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.

$$S(\theta) = \frac{d}{\theta} \log p(x|\theta). \tag{7}$$

スコア関数の $p(x|\theta)$ による平均は 0 である.

$$\int_{-\infty}^{\infty} \frac{d}{d\theta} \log p(x|\theta) p(x|\theta)\, dx \tag{8}$$

$$= \int_{-\infty}^{\infty} \frac{\frac{d}{d\theta} p(x|\theta)}{p(x|\theta)} p(x|\theta)\, dx \tag{9}$$

$$= \int_{-\infty}^{\infty} \frac{d}{d\theta} p(x|\theta)\, dx \tag{10}$$

$$= \frac{d}{d\theta} \int_{-\infty}^{\infty} p(x|\theta)\, dx \tag{11}$$

$$= 0. \tag{12}$$

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

$$I(\theta) = -\int_{-\infty}^{\infty} \left(\frac{d^2}{d\theta^2} \log p(x_i|\theta) \right) p(x|\theta)\, dx \tag{13}$$

と定義すると,

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

$$= -\int_{-\infty}^{\infty} \left(\left(\frac{\frac{d^2}{d\theta^2} p(x_i|\theta)}{p(x|\theta)} - \frac{(\frac{d}{d\theta} p(x_i|\theta))^2}{p(x|\theta)^2} \right) p(x|\theta) \right) dx \tag{15}$$

$$= -\int_{-\infty}^{\infty} \left(\frac{\frac{d^2}{d\theta^2} p(x_i|\theta)}{p(x|\theta)} \right) p(x|\theta)\, dx + \int_{-\infty}^{\infty} \left(\frac{(\frac{d}{d\theta} p(x_i|\theta))^2}{p(x|\theta)^2} \right) p(x|\theta)\, dx \tag{16}$$

スコア関数のときと同様, 第 1 項は消える. 第 2 項は

$$\int_{-\infty}^{\infty} \left(\frac{d}{d\theta} \log p(x|\theta) \right)^2 p(x|\theta)\, dx \tag{17}$$

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.

$$l_n(\theta) = \log\left(\frac{\prod_{i=1}^n p(x_i|\theta)}{\prod_{i=1}^n p(x_i|\theta_0)}\right) = \sum_{i=1}^n \log\left(\frac{\log p(x_i|\theta)}{\log p(x_i|\theta_0)}\right). \tag{18}$$

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,

$$\lim_{n\to\infty} l_n(\theta)/n = \int_{-\infty}^{\infty} p(x|\theta_0)\log\left(\frac{p(x|\theta)}{p(x_i|\theta)}\right) dx \tag{19}$$

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.

$$l_n(\theta) = \sum_{i=1}^n \log p(x|\theta)\, dx - \sum_{i=1}^n \log p(x_i|\theta_0)\, dx. \tag{20}$$

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,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \cdots \tag{21}$$

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,

$$l_n(\theta) = l_n(\theta_0) + l_n'(\theta_0)(\theta - \theta_0) + \frac{1}{2}l_n''(\theta_0)(\theta - \theta_0)^2 + \cdots \tag{22}$$

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

$$l_n(\theta_0 + h/\sqrt{n}) = l_n(\theta_0) + \frac{l_n'(\theta_0)}{\sqrt{n}}h + \frac{1}{2n}l_n''(\theta_0)h^2 + \cdots \tag{23}$$

[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. 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,

$$l_n(\theta_0 + h/\sqrt{n}) = \frac{l_n'(\theta_0)}{\sqrt{n}}h + \frac{1}{2n}l_n''(\theta_0)h^2 + O(1/\sqrt{n}) \tag{24}$$

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,

$$\begin{aligned} &l_n(\theta_0 + h/\sqrt{n}) \\ &\approx \frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n}\left(\frac{d}{d\theta}\log p(x_i|\theta)|_{\theta=\theta_0}\right)\right]h + \frac{1}{2n}\left[\sum_{i=1}^{n}\left(\frac{d^2}{d\theta^2}\log p(x_i|\theta)|_{\theta=\theta_0}\right)\right]h^2. \end{aligned} \tag{25}$$

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,

$$l_n(\theta_0 + h/\sqrt{n}) \approx Z\sqrt{I(\theta_0)}h - \frac{I(\theta_0)h^2}{2} \tag{26}$$

$$= -\frac{I(\theta_0)}{2}\left(h - \frac{Z}{\sqrt{I(\theta_0)}}\right)^2 + \frac{Z^2}{2} \tag{27}$$

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

$$l_n(\theta) = -\frac{I(\theta_0)}{2}\left(\sqrt{n}(\theta - \theta_0) - \frac{Z}{\sqrt{I(\theta_0)}}\right)^2 + \frac{Z^2}{2} \tag{28}$$

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.

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.

3　カイ 2 乗検定