最尤法とカイ二乗検定の基礎

黒木玄

2017年10月14日更新　(2017年10月14日作成)*

目 次

0　注意

すべての前提を正確に述べたり, 数学的に完璧に厳密な証明を書くつもりがないので読者は注意すること.

1　Kullback-Leibler 情報量と Fisher 情報量

1.1　Kullback-Leibler 情報量

$p(x)$, $q(x)$ は確率密度函数であるとする.
それらの **Kullback-Leibler 情報量** $D(q||p)$ を次のように定める:

$$D(q||p) = \int dx\, q(x) \log \frac{q(x)}{p(x)}.$$

*2017年10月14日: 作成.

Sanov の定理[1] より, KL 情報量 $D(q||p)$ は確率分布 p で確率分布 q をシミュレートしたときの「誤差」の大きさを表わす指標であるとみなせる[2]. KL 情報量が小さい方が確率分布 p による確率分布 q のシミュレーションの精度が高い.

だから, 未知の確率分布 q をより高い精度でシミュレートする確率分布 p を得るためには, KL 情報量 $D(q||p)$ がより小さくなると予想される確率分布 p を構成すればよい[3].

このノートでは KL 情報量の理論に深入りしないが, 読者の便のために以下の性質の証明の概略を説明しておく:

$$D(q||p) \geqq 0, \tag{1.1}$$

$$D(q||p) \geqq \frac{1}{2}\left(\int dx\ |q(x)-p(x)|\right)^2. \tag{1.2}$$

前者より後者の不等式の方が強い. 後者の強い方の不等式を使えば, Kullback-Leibler 情報量について $D(q||p)=0$ が成立していれば $p=q$ となることがわかる. 逆に $p=q$ ならば $\log(q(x)/p(x))=0$ なので $D(q||p)=0$ となることは自明である.

1.2 (1.1) の証明

$f(t)=t\log t$ と定める. $f'(t)=\log t+1$, $f''(t)=1/t$ なので $t>0$ で $f''(t)>0$ となり, $f(t)$ は $t>0$ で下に凸な函数であることがわかる. ゆえに Jensen の不等式[4] を確率分布 $p(x)$ に関する $f(q(x)/p(x))$ の平均に適用することによって,

$$D(q||p)=\int dx\ p(x)f\left(\frac{q(x)}{p(x)}\right) \geqq f\left(\int dx\ p(x)\frac{q(x)}{p(x)}\right)=f\left(\int dy\ q(x)\right)=f(1)=0.$$

Jensen の不等式に頼りたくなければ, $f(t)$ が下に凸な函数であることより $f(t)$ がその $t=1$ での接線 $t-1$ 以上であることを使えばよい. $f(t)\geqq t-1$ より

$$D(q||p)=\int dx\ p(x)f\left(\frac{q(x)}{p(x)}\right) \geqq \int dx\ p(x)\left(\frac{q(x)}{p(x)}-1\right)=\int dx\ (q(x)-p(x))=0.$$

1.3 (1.2) の証明

(1.1) よりも (1.2) の方がより強い不等式である. その分だけ証明は面倒になる.

x の積分領域を Ω と書く. Ω の部分領域 A に対して, $p(A)=\int_A dx\ p(x)$, $q(A)=\int_A dx\ q(x)$ のように書く. A の Ω における補集合を A^c と書く.

補題 1.1.

$$D(q||p) \geqq q(A)\log\frac{q(A)}{p(A)}+q(A^c)\log\frac{q(A^c)}{p(A^c)}$$

[1] Sanov の定理についてはこのノートでは解説しないが, 統計学を学ぶための必須の教養の一つだと思う. 統計学を学ぶために必要な確率論の基本的素養は, 大数の法則 (弱法則で十分), 中心極限定理, Sanov の定理 (大偏差原理) の 3 つである. Sanov の定理についてはよい解説が見付からなかったので, 自前で解説ノートを作った: https://genkuroki.github.io/documents/20160616KullbackLeibler.pdf

[2] p による q のシミュレーションと q による p のシミュレーションは異なるので, Kullback-Leibler 情報量が p,q について非対称であることは自然である.

[3] このような理由で未知の確率分布の推定を考える場合には Kullback-Leibler 情報量と Sanov の定理に関する知識が必須になる.

[4] Jensen の不等式の証明は https://genkuroki.github.io/documents/IntroProbability.pdf を参照せよ.

証明.

$$D(q||p) = \int_A dx\ q(x) \log \frac{q(x)}{p(x)} + \int_{A^c} q(x) \log \frac{q(x)}{p(x)}$$

より

$$\int_A dx\ q(x) \log \frac{q(x)}{p(x)} \geqq q(A) \log \frac{q(A)}{p(A)}$$

を証明すれば十分である. $f(X) = X \log X$ とおくと $f(X)$ は下に凸な函数になるのであった. A 上の確率密度函数 $p(x)/p(A)$ に関する平均に Jensen の不等式を適用すると

$$\begin{aligned}\int_A dx\ q(x) \log \frac{q(x)}{p(x)} &= \int_A dx\ p(x) f\left(\frac{q(x)}{p(x)}\right) = p(A) \int_A dx\ \frac{p(x)}{p(A)} f\left(\frac{q(x)}{p(x)}\right) \\ &\geqq p(A) f\left(\int_A dx\ \frac{p(x)}{p(A)} \frac{q(x)}{p(x)}\right) \\ &= p(A) f\left(\frac{q(A)}{p(A)}\right) = q(A) \log \frac{q(A)}{p(A)}. \quad \square\end{aligned}$$

補題 1.2. a, b が 0 と 1 のあいだの実数のとき

$$b \log \frac{b}{a} + (1-b) \log \frac{1-b}{1-a} \geqq 2(b-a)^2$$

証明. 不等式の左辺から右辺を引いた結果を $f(a)$ と書く:

$$f(a) = b \log \frac{b}{a} + (1-b) \log \frac{1-b}{1-a} - 2(b-a)^2.$$

$f(a)$ の最小値が 0 であることを示せば十分である. このとき,

$$f'(a) = -\frac{b}{a} + \frac{1-b}{1-a} - 4(a-b) = \frac{a-b}{a(1-a)} - 4(a-b) = (a-b)\left(\frac{1}{a(1-a)} - 4\right).$$

$a(1-a)$ の最大値は $1/4$ なのでその逆数の最小値は 4 である. ゆえに $f'(a)$ の符号は $a-b$ の符号に等しい. $a < b$ で $f(a)$ は単調減少し, $a > b$ で $f(a)$ は単調増加する. ゆえに $a = b$ で $f(a)$ は最小になる. その最小値は $f(b) = 0$. $\square$

補題 1.3. $A = \{\, x \in \Omega \mid p(x) \leqq q(x) \,\}$ とおくと

$$\int dx\ |q(x) - p(x)| = 2(q(A) - p(A)).$$

証明.

$$\begin{aligned}\int dx\ |q(x) - p(x)| &= \int_A dx\ (q(x) - p(x)) + \int_{A^c} dx\ (p(x) - q(x)) \\ &= q(A) - p(A) + p(A^c) - q(A^c) \\ &= q(A) - p(A) + (1 - p(A)) - (1 - q(A)) = 2(q(A) - p(A)). \quad \square\end{aligned}$$

以上の 2 つの補題を使って, (1.2) を証明しよう.
$A = \{ x \in \Omega \mid p(x) \leqq q(x) \}$ とおき, $a = p(A)$, $b = q(A)$ とおく. 補題 1.1 より

$$D(q||p) \geqq b \log \frac{b}{a} + (1-b) \log \frac{1-b}{1-a}.$$

補題 1.2 より

$$b \log \frac{b}{a} + (1-b) \log \frac{1-b}{1-a} \geqq 2(b-a)^2.$$

補題 1.3 より

$$2(b-a)^2 = \frac{1}{2}(2(b-a))^2 \geqq \frac{1}{2} \left(\int dx \, |q(x) - p(x)| \right)^2.$$

以上を合わせれば (1.2) が得られる.

1.4 Fisher 情報量

$p_w(x) = p(x|w)$ は r 個のパラメーター $w = (w_1, \ldots, w_r)$ を持つ x に関する確率密度函数であるとし, $h = (h_1, \ldots, h_r)$ の函数

$$D(h) = D(p_w||p_{w+h}) = \int dx \, p(x|w) \log \frac{p(x|w)}{p(x|w+h)}$$

について考えよう.
まず $D(0) = D(p_w||p_w) = 0$ であり, Kullback-Leibler 情報量が 0 以上であることより, $D(h) \geqq 0$ である. ゆえに $D(h)$ は $h = 0$ で極小になるので, $\partial D(0)/\partial h_i = 0$ となる. そうなることは次のようにして, 直接的にも確かめられる:

$$\begin{aligned} \frac{\partial D(0)}{\partial h_i} &= -\int dx \, p(x|w) \frac{\partial \log p(x|w)}{\partial w_i} = -\int dx \, \frac{\partial p(x|w)}{\partial w_i} \\ &= -\frac{\partial}{\partial w_i} \int dx \, p(x|w) = -\frac{\partial}{\partial w_i} 1 = 0. \end{aligned}$$

これは, 確率分布 $p(x|w)$ に関する $\partial \log p(x|w)/\partial w_i$ の平均が 0 であることも意味する.
以上によって $D(h)$ の $h = 0$ での Taylor 展開の 1 次以下の項が消えることがわかった. 本質的に 2 次の項が Fisher 情報量である. Fisher 情報量がわかると h が小さなときの KL 情報量 $D(h)$ の様子がわかる.
次のようにして定義される実対称行列 $I(w) = [I_{ij}(w)]_{i,j=1}^r$ を **Fisher 情報量**と呼ぶ:

$$I_{ij}(w) = \frac{\partial^2 D(0)}{\partial h_i \partial h_j} = -\int dx \, p(x|w) \frac{\partial^2 \log p(x|w)}{\partial w_i \partial w_j}$$

$D(h)$ は $h = 0$ で極小になるので, Fisher 情報量 I の固有値はすべて 0 以上になる [5].

[5] 実対称行列に関する線形代数. 線形代数はとても大事. おそらく, 大学 1 年で線形代数について習っただけだと, 線形代数の習得は不十分である. 大学 4 年間かけないと十分に習得できない可能性が高い.

Fisher 情報量は次のようにも表せる:

$$I_{ij}(w) = \int dx\ p(x|w)\ \frac{\partial p(x|w)}{\partial w_i}\ \frac{\partial p(x|w)}{\partial w_j}.$$

この公式は, Fisher 情報量 $I(w) = [I_{ij}(w)]$ は確率分布 $p(x|w)$ に関する $\partial \log p(x|w)/\partial w_i$ の分散共分散行列であることを意味している. その公式を証明しよう. $I_{ij}(w)$ の定義より

$$\begin{aligned} I_{ij} &= -\int dx\ p(x|w) \left(\frac{\partial^2 p(x|w)/\partial w_i \partial w_j}{p(x|w)} - \frac{\partial p(x|w)/\partial w_i\ \partial p(x|w)/\partial w_j}{p(x|w)^2} \right) \\ &= -\int dx\ \frac{\partial^2 p(x|w)}{\partial w_i \partial w_j} + \int dx\ p(x|w)\ \frac{\partial p(x|w)}{\partial w_i}\ \frac{\partial p(x|w)}{\partial w_j} \end{aligned}$$

であり,

$$\int dx\ \frac{\partial^2 p(x|w)}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \int dx\ p(x|w) = 0$$

となるので, 上の公式が成立している.

2　最尤法と Wilks の定理とカイ二乗検定

第 1.4 節の設定をそのまま用いる.

この節では, $w = 0$ での Fisher 情報量 $I(0) = [I_{ij}(0)]$ の固有値はすべて正であると仮定する [6].

2.1　対数尤度比の漸近挙動

$X, X_1, X_2, \ldots$ は確率分布 $p(x|0)$ に従う独立同分布確率変数列であるとする [7].

-1 倍された対数尤度比函数 $\ell_n(w)$ を次のように定める:

$$\ell_n(w) = -\log \frac{\prod_{k=1}^n p(X_k|w)}{\prod_{k=1}^n p(X_k|0)} = \sum_{k=1}^n \log \frac{p(X_k|0)}{p(X_k|w)}.$$

パラメーター w の函数 $L_n(w) = \prod_{k=1}^n p(X_k|w)$ をサイズ n のサンプル $X_1, X_2, \ldots, X_n$ の**尤度函数**と呼ぶ. $\ell_n(w)$ は尤度の比の対数の -1 倍である. $\ell_n(w)$ を最小化する $w = \hat{w}$ は尤度 $L_n(w)$ を最大化する.

大数の法則より,

$$\lim_{n\to\infty} \frac{1}{n} \ell_n(w) = \int dx\ p(x|0) \log \frac{p(x|0)}{p(x|w)} = D(p_0||p_w).$$

---

[6] 正則性の仮定. 現実への応用の場合には「精度的に 0 とみなすことが妥当なほど微小な固有値を $I(0)$ が持たない」という条件で置き換える必要がある. この正則性の条件は複雑な $p(x|w)$ については成立していない場合がかなり多い.

[7] 以下で使う確率論の結果については https://genkuroki.github.io/documents/IntroProbability.pdf を参照. この解説では中心極限定理の特性函数やモーメント母函数を使わない証明が紹介されている. 本質的に Taylor の定理しか使っていない.

これより, $\ell_n(w)$ を最小化する $\hat{w}$ に対応する確率分布 $p(x|\hat{w})$ は $p(x|0)$ を近似していると予想される.

そのことを確認するために n で割る前の $\ell_n(w)$ の $n \to \infty$ での様子を調べたい. そのためには $\ell_n(w)$ に $w = 0 + h/\sqrt{n}$ を代入して, $h = 0$ で Taylor 展開すればよい. 自明に $\ell_n(0) = 0$ が成立していることがわかるので, その Taylor 展開は次のような形になる:

$$\begin{aligned}\ell_n\left(\frac{h}{\sqrt{n}}\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{r}\frac{\partial \ell(0)}{\partial w_i}h_i + \frac{1}{2n}\sum_{i,j=1}^{r}\frac{\partial^2 \ell(0)}{\partial w_i \partial w_j}h_i h_j + \cdots \\ &= -\sum_{i=1}^{r}\left[\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{\partial \log p(X_k|0)}{\partial w_i}\right]h_i - \frac{1}{2}\sum_{i,j=1}^{r}\left[\frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2 \log p(X_k|0)}{\partial w_i \partial w_j}\right]h_i h_j + \cdots\end{aligned}$$

$\partial \log p(X|0)/\partial w_i$ の平均は 0 であり, その分散共分散行列は Fisher 情報量 $I(0) = [I_{ij}(0)]$ に一致する. さらに $\partial^2 p(X|0)/\partial w_i \partial w_j$ の平均の -1 倍は $I_{ij}(0)$ の定義そのものである. ゆえに中心極限定理と大数の法則より, 分散共分散行列が $I(0)$ の多変量正規分布に従う確率変数 $(W_1, \ldots, W_r)$ が存在して, $n \to \infty$ で

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{\partial \log p(X_k|0)}{\partial w_i} \to W_i, \qquad \frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2 \log p(X_k|0)}{\partial w_i \partial w_j} \to -I_{ij}(0).$$

大数の法則より, "$+\cdots$" の部分は $O(1/\sqrt{n})$ で 0 に収束することもわかる. これで $n \to \infty$ で

$$\ell_n\left(\frac{h}{\sqrt{n}}\right) = -\sum_{i=1}^{r}W_i h_i + \frac{1}{2}\sum_{i,j=1}^{r}I_{ij}(0)h_i h_j + o(1)$$

となることがわかった.

変数 h_i 達を線形座標変換して, さらにすっきりした形で表示しよう.

$I(0) = [I_{ij}(0)]$ は固有値がすべて正の実対称行列なので, 固有値がすべて正の実対称行列 Σ が存在して

$$I(0) = \Sigma^2$$

と表せる. h_i 達を成分に持つ列ベクトルを h と書き, W_i 達を成分に持つ列ベクトルを W_i と書くことにする. W_i 達の分散共分散が $I(0) = \Sigma^2$ になることは

$$E[WW^T] = I(0) = \Sigma^2$$

となることを意味する. ここで $E[\ ]$ は期待値を取る操作で, $(\ )^T$ は転置を取る操作である. 両辺を左右から Σ^{-1} をかけると,

$$E[(\Sigma^{-1}W)(\Sigma^{-1}W)^T] = E$$

を得る. これは Z_i を成分に持つ列ベクトルを $Z = \Sigma^{-1}W$ と定めると Z_i 達の分散共分散行列が単位行列 E になることを意味している. Z_i 達は標準正規分布に従う独立な確率変

数達になる. ξ_i を成分に持つ列ベクトル ξ を $\xi = h$ と定めると,

$$\begin{aligned}-\sum_{i=1}^{r} W_i h_i + \frac{1}{2}\sum_{i,j=1}^{r} I_{ij}(0) h_i h_j &= -W^T h + \frac{1}{2} h^T I(0) h = -Z^T \Sigma h + \frac{1}{2} h^T \Sigma^T \Sigma h \\ &= -Z^T \xi + \frac{1}{2}\xi^T \xi = \frac{1}{2}(\xi - Z)^T (\xi - Z) - \frac{1}{2} Z^T Z \\ &= \frac{1}{2}\sum_{i=1}^{r} (\xi_i - Z_i)^2 - \frac{1}{2}\sum_{i=1}^{r} Z_i^2.\end{aligned}$$

2.2 最尤法とカイ二乗検定の基礎

前節の結果をまとめると次を得る:

$$\ell_n\left(\frac{h}{\sqrt{n}}\right) = \frac{1}{2}\sum_{i=1}^{r}(\xi_i - Z_i)^2 - \frac{1}{2}\sum_{i=1}^{r} Z_i^2 + O\left(\frac{1}{\sqrt{n}}\right), \quad h = \Sigma^{-1}\xi, \quad \Sigma = \sqrt{I(0)}.$$

ゆえに $\ell_n(h/\sqrt{n})$ を最小化する $h = h^*$ は

$$h^* \approx \Sigma^{-1} Z$$

を満たしており, $\Sigma^{-1}Z$ は分散共分散行列が $\Sigma^{-2} = I(0)^{-1}$ の多変量正規分布に従う:

$$h^* \approx \Sigma^{-1} Z \sim \mathrm{Normal}(0, I(0)^{-1}).$$

ゆえに, $w^* = h^*/\sqrt{n}$ については

$$w^* \approx \frac{1}{\sqrt{n}}\Sigma^{-1} Z \sim \mathrm{Normal}(0, I(0)^{-1}/n).$$

w^* の分散共分散行列は $O(1/n)$ のオーダーで減少する.

$\ell_n(w)$ は尤度比の対数の -1 倍であった. **ゆえに, 尤度函数 $L_n(h/\sqrt{n})$ を最大にする $h = h^*$ は分散共分散行列が Fisher 情報量の逆行列であるような多変量正規分布に近似的に従う確率変数になる**[8]. これが**最尤法**の基礎である.

w_0 の十分小さな近傍を H_1 とすると,

$$\min_{w \in H_1} \ell_n(w) \approx -\frac{1}{2}\sum_{i=1}^{r} Z_i^2$$

を満たしている. この右辺の -2 倍は自由度 r のカイ二乗分布に従う.

$\ell_n(w)$ は対数尤度比の -1 倍だったので

$$2\log\frac{\max_{w\in H_1} L_n(w)}{L_n(0)} \approx \sum_{i=1}^{r} Z_i^2 \sim \chi^2(r).$$

すなわち, **最大尤度と尤度 $L_n(0)$ の比の対数の 2 倍は自由度が全パラメーター数 r に等しいカイ二乗分布に近似的に従う.**

[8] もしも $I(0)$ のある固有値が微小ならば h^* のある成分の分散は巨大になってしまう.

上の結果はさらに以下のように一般化される. $h_{s+1} = h_{s+2} = \cdots = h_r = 0$ で定義される H_1 の部分集合を H_0 と書くとき, 必要なら座標系 ξ をさらに回転させることによって, $\xi_{s+1} = \xi_{s+2} = \cdots = \xi_r = 0$ で定義される部分集合が H_0 に一致するようにできる. そのとき

$$2\log\frac{\max_{w\in H_0} L_n(w)}{L_n(0)} \approx \sum_{i=1}^{s} Z_i^2 \sim \chi^2(s)$$

なので, 上の結果との差を取ることによって,

$$2\log\frac{\max_{w\in H_0} L_n(w)}{\max_{w'\in H_0} L_n(w')} \approx \sum_{i=s+1}^{r} Z_i^2 \sim \chi^2(r-s).$$

すなわち, r 次元の H_1 における最大尤度と H_1 の s 次元の部分集合 H_0 における最大尤度の比の対数の 2 倍は近似的に自由度 $r-s$ のカイ二乗分布に従う.

この結果を **Wilks の定理**と呼ぶ. Wilks の定理はカイ二乗検定の基礎である.

通常, r 次元の H_1 の s 次元の部分集合 H_0 として, 帰無仮説に対応するパラメーターの集合が採用される. もしも, サンプルを生成した真の確率分布が H_0 に入っているならば, 最尤法で求めた数値的な最大尤度比の対数の 2 倍

$$\chi^2 = 2\log\frac{\max_{w\in H_1} L_n(w)}{\max_{w'\in H_0} L_n(w')}$$

は自由度 $r-s$ のカイ二乗分布に近似的に従っているはずなので, 自由度 $r-s$ のカイ二乗分布にしてはその値が大き過ぎるということになるならば, これがカイ二乗検定の一般的な仕組みである.

2.3 赤池情報量基準 AIC への応用

パラメーター数が r の場合の赤池情報量基準 AIC_1 は

$$\mathrm{AIC}_1 = -2\log\max_{w\in H_1} L_n(w) + 2r$$

と定義される. 同様に

$$\mathrm{AIC}_0 = -2\log\max_{w\in H_0} L_n(w) + 2s.$$

AIC は小さい方が, 予測分布の予測精度が高いと推定される. 上の 2 つの AIC について

$$\mathrm{AIC}_1 < \mathrm{AIC}_0$$

が成立するための必要十分条件は

$$2\log\frac{\max_{w\in H_1} L_n(w)}{\max_{w'\in H_0} L_n(w')} > 2(r-s).$$

サンプルを生成した真の分布のパラメーターが H_0 に含まれているならば, n が大きいとき, 左辺は自由度 $r-s$ のカイ二乗分布に近似的に従う. それにも関わらず, H_1 の AIC の

方が H_0 の AIC よりも小さくなってしまう確率の近似値 $P(\chi^2(r-s) > 2(r-s))$ は次のようになる:

$$\begin{aligned}
P(\chi^2(1) > 2) &= 15.73\%, \\
P(\chi^2(2) > 4) &= 13.53\%, \\
P(\chi^2(3) > 6) &= 11.16\%, \\
P(\chi^2(4) > 8) &= 9.16\%.
\end{aligned}$$

これらの数値は AIC が真のモデルの選択に失敗する確率である. サンプルサイズがどんなに大きくなっても AIC によるモデル選択はこの確率で真のモデル選択に失敗する. 5%の有意水準を満足させるためには

$$\begin{aligned}
P(\chi^2(1) > 3.84) &= P(\mathrm{AIC}_0 - \mathrm{AIC}_1 > 1.84) = 5\%, \\
P(\chi^2(2) > 5.99) &= P(\mathrm{AIC}_0 - \mathrm{AIC}_1 > 1.99) = 5\%, \\
P(\chi^2(3) > 7.81) &= P(\mathrm{AIC}_0 - \mathrm{AIC}_1 > 1.81) = 5\%, \\
P(\chi^2(4) > 9.49) &= P(\mathrm{AIC}_0 - \mathrm{AIC}_1 > 1.49) = 5\%.
\end{aligned}$$

1%の有意水準では

$$\begin{aligned}
P(\chi^2(1) > 6.63) &= P(\mathrm{AIC}_0 - \mathrm{AIC}_1 > 4.63) = 1\%, \\
P(\chi^2(2) > 9.21) &= P(\mathrm{AIC}_0 - \mathrm{AIC}_1 > 5.21) = 1\%, \\
P(\chi^2(3) > 11.34) &= P(\mathrm{AIC}_0 - \mathrm{AIC}_1 > 5.34) = 1\%, \\
P(\chi^2(4) > 13.28) &= P(\mathrm{AIC}_0 - \mathrm{AIC}_1 > 5.28) = 1\%.
\end{aligned}$$

最大尤度比の対数の 2 倍の値 (もしくは AIC の値) でモデル選択する場合には以上の数値を覚えておいた方がよいだろう.

注意 2.1 (警告). 最大尤度比の対数の 2 倍によるカイ二乗検定を使用する場合には, 比較する確率モデルのあいだに $H_1 \supset H_0$ (パラメーターの集合の次元が下がる) という包含関係が無ければいけない. □

注意 2.2. 以上では $H_1 \supset H_0$ の場合のモデルの比較を扱ったが, AIC による比較はそのような包含関係が無くても可能である.

ただし, AIC によるモデル選択は予測分布の予測精度の推定値の比較によるモデル選択であり, 真のモデルを選択することではない. 同様のことが WAIC にも言える. 真のモデルの選択が目的であれば BIC または WBIC を使う方が適切である. □