私のための「頻度論」統計学入門

2020 年 2 月 11 日

目 次

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. いまそれらを勉強しているところなので, 後で書く.

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) 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）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 $i = 1,\ldots,M$ 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.

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 $D_{\mathrm{KL}}(p\|q) = D_{\mathrm{KL}}(q\|p)$ は成り立たない.

1.4　シミュレーション

サンプルサイズが増えれば増えるほど正規分布でないことがはっきりしてくる.

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.

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

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.

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

これをサンプルサイズ（標本の大きさ） n で割ると大数の法則により,

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

[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.

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

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,

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

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

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

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,

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

である. ここで, $E_g[X] = \int xg(x)\,dx$.

（途中計算は省略して）

$\hat{\mu} = 4, \quad \hat{\sigma}^2 = 4$ 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.

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

また $Z_1, Z_2, \ldots, Z_r$ が独立でそれぞれ標準正規分布に従うとき, $Z_1^2 + Z_2^2 + \cdots + Z_r^2$ 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そうなことがわかる.

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,

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

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.

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

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

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

ここで $\hat{q}_i/\hat{q}_0 = a_{i1}/(N_iM_i/N)$, $(1-\hat{q}_i)/(1-\hat{q}_0) = a_{i1}/(N_iM_i/N)$ 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_iM_i/N や N_iM_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.

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

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	$E_{11} = N_1M_1/N$	$E_{12} = N_1M_2$	N_1
暴露なし	$E_{21} = M_1N_2/N$	$E_{22} = M_2N_2/N$	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.

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

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

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.

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

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

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

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

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

$\Lambda = (\lambda_{ij})$ を正の実数を成分とする 2×2 行列とする. a_{ij} がそれぞれ独立にパラメータ λ_{ij} のポアソン分布に従うとすると, A が生じる確率 $p(A|\Lambda)$ は,

$$p(A|\Lambda) = \prod_{i,j}\frac{e^{-\lambda_{ij}}\lambda_{ij}^{a_{ij}}}{a_{ij}!} \tag{48}$$

と書ける. 分割表の各行と各列は独立であるとするモデルについては表 2 との対応を考え, 行についてのパラメータ μ と列についてのパラメータ ν, 観測の総数を規定するパラメータ τ を用いて,

注意点. 統計的仮説検定は確率版背理法と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 4.1 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のとき, 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.

本当は分割表の各行と各列が独立であるときに, 分割表の各行と各列が独立でないと誤って判断する確率を α エラーと呼ぶ. 本当は分割表の各行と各列が独立でないときに, 分割表の各行と各列が独立でないと正しく判断する確率を検出力（power）と呼ぶ.

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いいように思う. 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いいように思う. 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

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 $(i = 1, \ldots, n)$ 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.

$$I_o(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p(x_i|\theta). \tag{49}$$

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.

θ が不明なので, 最尤推定量 $\hat{\theta}$ を用いて近似するわけである. このように計算していい根拠は実はよくわからない. ここでいうわからないは, 私個人がわかっていないという意味であり, 統計学的に未解決ということではない.

上記の性質から, ワルド検定統計量 z_0 を次のように定義する.

$$z_0 = \frac{\hat{\theta} - \theta_0}{\mathrm{se}(\hat{\theta})}, \tag{50}$$

ここで $\mathrm{se}(\hat{\theta}) = (nI_o(\theta))^{-1/2}$ とおいた.

カイ二乗検定の説明ではかなり省略して書いたので, ここで仮説検定の枠組みを説明しなおす. 仮説検定では次のように「仮説」を設定する.

$$H_0 : \ \theta = \theta_0 \tag{51}$$

$$H_1 : \ \theta \neq \theta_0 \tag{52}$$

$$\tag{53}$$

H_0 を帰無仮説, H_1 を対立仮説という. (57) 式での θ_0 は帰無仮説 $H_0 : \theta = \theta_0$ と対応している. 求めた検定統計量 z_0 より絶対値の大きい値が

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

出る確率（p 値）を標準正規分布から求め, p 値があらかじめ定めた有意水準より小さいとき, 帰無仮説が棄却されたという. この形の検定を両側検定という.

一方で仮説を次のように設定することもある.

$$H_0 : \theta = \theta_0 \tag{54}$$

$$H_1 : \theta > \theta_0 \tag{55}$$

$$\tag{56}$$

この場合, z_0 より大きい値が出る確率（p 値）を標準正規分布から求め, p 値があらかじめ定めた有意水準より小さいとき, 帰無仮説が棄却されたという. この形の検定を片側検定という.

片側検定の場合, 帰無仮説を $\theta \leq \theta_0$ と書きたくなるかもしれないが, 仮説検定では帰無分布（帰無仮説のもとで検定統計量の従う分布）が 1 つに定まらないと, 計算ができないので, $H_0 : \theta = \theta_0$ という書き方で正しい.

対立仮説を $H_1 : \theta < \theta_0$ と設定しても構わない. その場合は z_0 より小さい値が出る確率を p 値とする.

注意点. 統計的仮説検定は確率版背理法と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. 帰無仮説が棄却された場合であっても, $\theta = \theta_0$ か $\theta \neq \theta_0$ かという, はっきりとした 2 値的な判断をすることは難しい.

統計モデルがサンプルを生成した分布からかけ離れていた場合にも, 帰無仮説は棄却されやすくなる. 例えば独立同分布の仮定というのも, 統計モデリングの一環である. 統計モデルにおける独立同分布は, 現実の世界でのランダムサンプリングに対応すると思われる.

4.2 ワルド信頼区間

$100(1-\alpha)\%$ ワルド信頼区間を次のように定義する.

$$[\hat{\theta} - \mathrm{se}(\hat{\theta}) z_{\alpha/2}, \hat{\theta} + \mathrm{se}(\hat{\theta}) z_{\alpha/2}] \tag{57}$$

これはワルド検定において, θ_0 を動かしたとき, 有意水準 α で帰無仮説が棄却されない範囲を求めている.

統計モデルによってデータを生成した分布が実現可能であり, サンプルサイズが十分に大きいとき, ワルド信頼区間は次の性質を満たすことが期待される.

$$\Pr[\hat{\theta} - \mathrm{se}(\hat{\theta}) z_{\alpha/2} \leq \theta \leq \hat{\theta} + \mathrm{se}(\hat{\theta}) z_{\alpha/2}] \geq 1 - \alpha. \tag{58}$$

伝統的な統計学の教科書では, この性質

$$\Pr[T_1 \leq \theta \leq T_2] \geq 1 - \alpha \tag{59}$$

を満たす $[T_1, T_2]$ を信頼区間と呼ぶ, と定義している場合もある. しかし, 信頼区間の作り方は統計モデルに依存するため, 現実の場合, 求めた信頼区間が上記の性質をみたしているかはわからない.

そこで, 両側検定を行ったときに帰無仮説 $\theta = \theta_0$ が棄却されない θ_0 の範囲のことを信頼区間と呼ぶ, と定義するほうが明快だと思う.

統計モデルによってデータを生成した分布が実現可能なときの $\Pr[T_1 \leq \theta \leq T_2]$ を被覆確率（covarage probability）という. 被覆確率を計算するためには, コンピューターを使ってシミュレーションを行うのが一般的である.

また信頼区間について, 多くの伝統的な統計学の教科書では, 分布しているのは θ ではなく $[T_1, T_2]$ だから, θ が $[T_1, T_2]$ に含まれる確率が $1 - \alpha$ であると言ってはいけない, と強調している .

ではどのように解釈するのかというと, 何度もデータを取り, 信頼区間を作ることを繰り返したとき, 信頼区間が θ を含んでいる割合が $1 - \alpha$ だと述べられる.

しかし, 「信頼区間が θ を含んでいる」は「θ が信頼区間に含まれる」と同じ意味であるし, この場合の「割合」は「確率」の言い換えに過ぎない. したがって, このような強調が本質的であるとは私には思えない.

5　回帰型の統計モデル

書くかどうか迷っている. これについては久保（2012）などを読んだほうがいいかもしれない.

5.1　ロジスティック回帰

回帰と分類の違いを丁寧に説明する資料もあるが, 私にはそのような区別が本質的であるとは思えない.

注意点. 推定されたパタメータの解釈は意外と難しい部分がある. 推定されたパラメータが統計的に「有意」になっていたとしても, 統計モデルを少し変えるだけでまったく別の結果が出るということは, 頻繁に起こる.

6 モデル選択

統計学の解説では, 正規分布ならば t 検定, 正規分布でないならばノンパラメトリック検定, ラベルデータが与えらていれば識別モデルを使い, 与えらていなければクラスタリング, ……といった, 複雑なフローチャートを記載してあるものがある.

このようなフローチャートは現実的にはあまり役に立たないことが多い. 例えば「正規分布ならば t 検定」といっても, データが正規分布であるかどうかはわからない[4].

本稿は統計モデルは目的に合わせて自由に選ぶという立場をとる. しかし, 選んだモデルが良いかどうかは別の問題である. そこで, 作った統計モデルを評価する方法が重要になる.

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している.

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になる.

[4]むしろ, 現実のデータが都合よく正規分布になっていることはありえないと断言したほうがいいかもしれない

7　いくつかの統計モデルのカタログ

統計モデルの紹介がされている楽しい本に松浦（2016）や須山（2017）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．

7.1　幾何分布回帰

あるとき，とあるウェブページのアクセス記録のデータを受け取ったとしよう．その一部を抜粋したものが，表 3 である．

表 3: pageDepth データ

pageDepth	userGender	userAgeBracket	userType	sessions
1	female	18-24	New Visitor	2996
1	female	18-24	Returning Visitor	382
1	female	25-34	New Visitor	3426
1	female	25-34	Returning Visitor	474
1	female	35-44	New Visitor	1850
1	female	35-44	Returning Visitor	218

各列の意味を聞いてみると，pageDepth はそのウェブページから離脱するまでに閲覧したページの数，userGender は性別，userAgeBracket は年齢層，userType は新規訪問者かリピーターかを表し，sessions は訪問ののべ人数だという．

このデータからウェブページをたくさん閲覧しやすい（回遊しやすい）のはどのユーザー層か知りたいとする．

pageDepth を目的変数として回帰を行えば，このデータセットをうまく要約できそうな気がする．

研究室の後輩に「回帰ってなにがあるんですか？　ポアソン回帰とかロジスティック回帰とか、あとは」と聞かれたことがある．その答えはもちろん「いくらでもある」である[5]．そのモデルが良いかどうかは別の問題として，好きな分布を使って自由に回帰をしてよい．

[5]おそらく，このように質問させた事自体が，私のそれまでの指導がまずかったことを意味している．

このようなデータに対してはどんな分布を用いるのが適切だろうか.

現象を思い切り単純化して考え, コインの表が出たらページを離脱し, コインの裏が出たらページを離脱するとする. このとき, 表が出る確率を p とすると, はじめて表が出るまでのコイントスの回数 y の分布は,

$$f(y|p) = (1-p)^{y-1}p \tag{60}$$

で表せる. このような分布を幾何分布と呼ぶ. 幸か不幸か, R の glm には幾何分布を使った回帰が実装されていないようなので, ここでは手作りで幾何分布回帰を行う.

本解析の場合 y は pageDepth に対応する. p は 0 から 1 の範囲の値を取るパラメータなので, ロジスティック回帰のときと同様, 逆ロジット変換を利用して,

$$p_n = \frac{1}{1+\exp(-X_n\beta)} \tag{61}$$

という形で, 回帰構造を持たせることにする.

関数の最小値, または最大値を求めることを最適化という. R には optim という最適化用の関数があり, これに対数尤度関数を与えれば, 簡単に最尤法を実行することができる[6].

R では, 対数尤度関数は次のように書けばよい.

```
ll <- function(beta,y,X,w){
  logprob <- plogis(X %*% beta,log.p = TRUE)
  logprob2 <- plogis(-X %*% beta,log.p = TRUE)
  -sum(w*((y-1)*logprob2+logprob))
}
```

R の optim はデフォルトでは最小値を求めるため, 対数尤度に -1 を書けた値を返すように, 関数を定義した.

optim は次のように使う.

```
optim(ini, fn = ll,
      method = "BFGS",
      y=y,X=X,w=w, hessian = TRUE)
```

[6]ただし, 最適化はいつもうまくいくとは限らない. 関数に極値がいくつもあるような場合, 最適化はたやすく失敗する

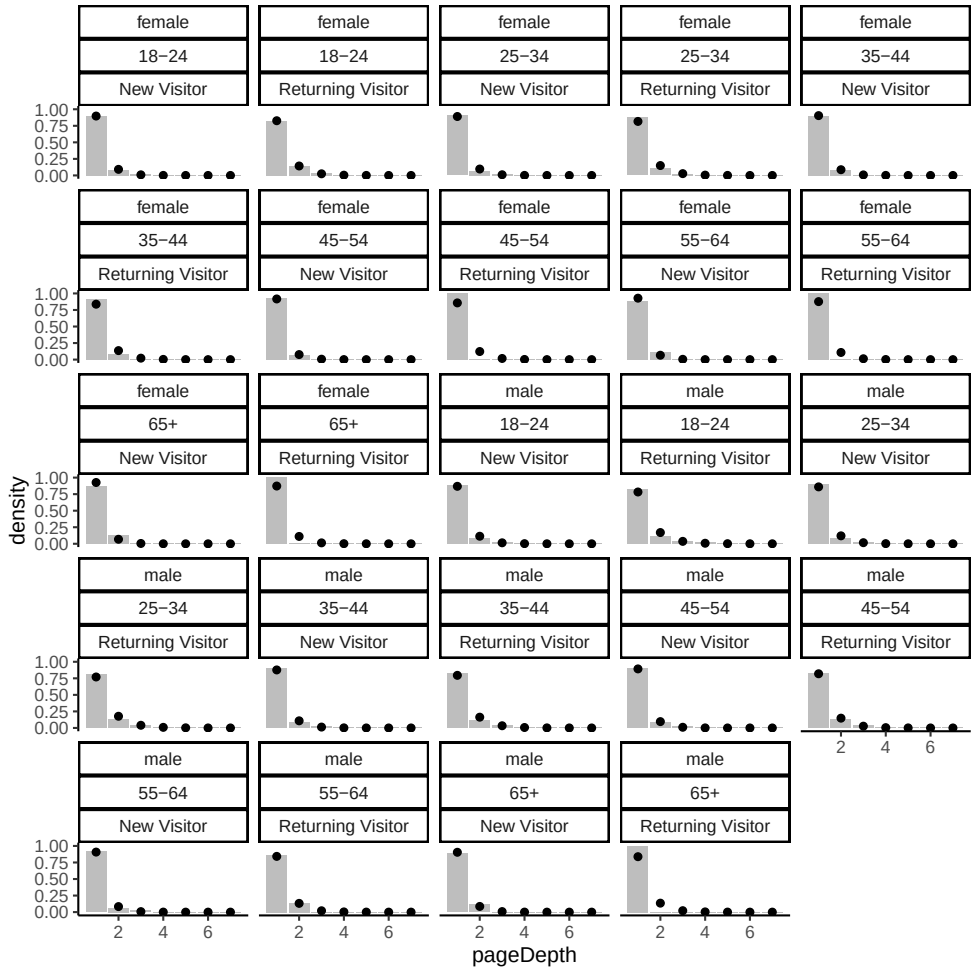

図 1: 当てはまりの確認. 棒グラフ: 観測値, 点: 推定値.

ここで `ini` はパラメータの初期値である. `hessian = TRUE` とすることで, 数値微分による目的関数のヘシアンが求まる. これはサンプルサイズで割る前の観測情報量そのものである.

当てはめた幾何分布と, 実際のデータを並べてプロットしてみよう（図 1）. 大きな計算ミスはなさそうである.

係数 β の推定値と, 観測情報量より計算した 2 標準誤差のエラーバーをプロットしてみよう（図 2）.

リピーターであるか否かが, サイトを回遊しやすいかどうかに大きく寄与する. また, 25-34 歳の男性がサイトを回遊しやすい傾向がわかる.

コードの全体は, R フォルダにある geomreg.R ファイルである.

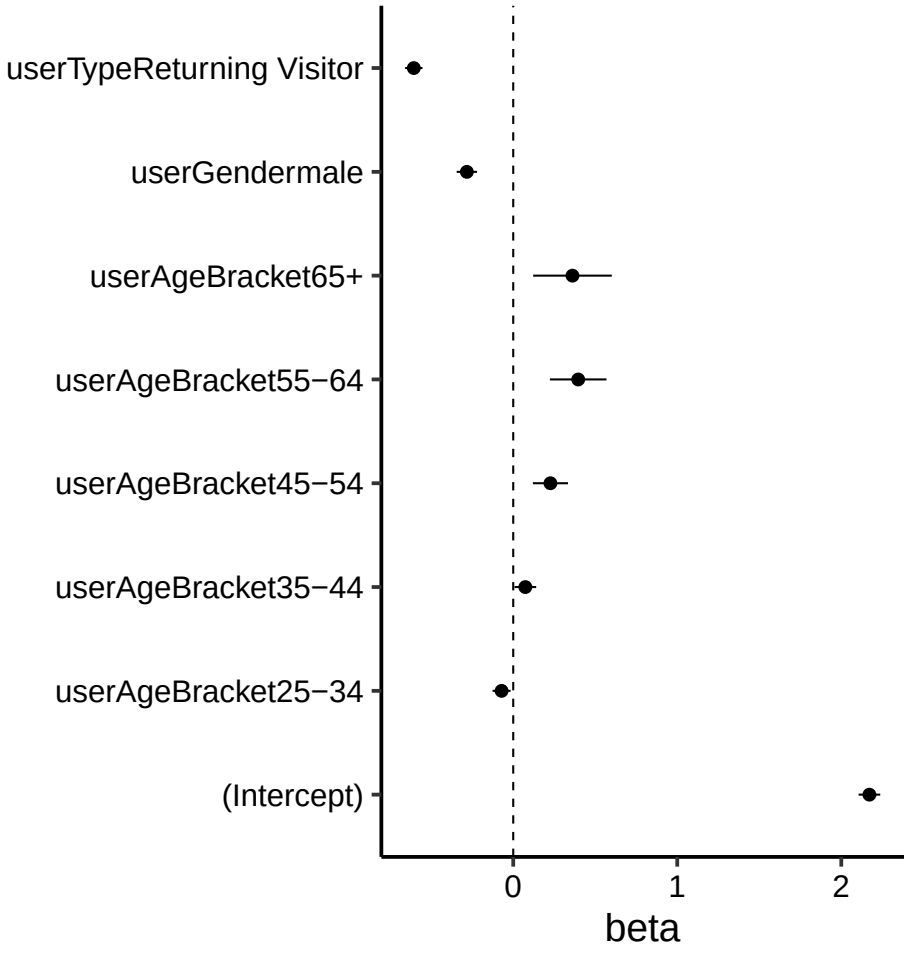

図 2: 係数のプロット. バーは 2 標準誤差.

7.2　生存時間分析

幸か不幸か, R の survreg にはガンマ分布を使った回帰が実装されていないようなので, ここでは手作りでガンマ回帰を行う.

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, $\sigma = \exp(\rho)$ とおいて, ρ を最尤推定し, $\hat{\sigma} = \exp(\hat{\rho})$ として構わない.

7.3　比例ハザードモデル

7.4　非定常ポアソン過程

時系列データというと, 等間隔の時点ごとに計測されたデータを思い浮かべることが多い. しかし, イベントの生起がランダムで, イベントの起こった時点だけが記録されていることもある.

参考文献

[1] 黒木玄. (2017). 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.

[3] 久保拓弥. (2012). データ解析のための統計モデリング入門—般化線形モデル・階層ベイズモデル・MCMC. 岩波書店.

[4] 松浦健太郎. (2016) Stan と R でベイズ統計モデリング. 共立出版.

[5] 須山敦史. (2017) ベイズ推論による機械学習入門. 講談社.