

信頼区間・信用区間・最高事後密度領域

樋口さぶろお <https://hig3.net>

龍谷大学大学院 先端理工学部研究科 数理・情報科学コース

計算科学特論 L04(2024-10-16 Wed)

最終更新: Time-stamp: "2024-10-16 Wed 12:21 JST hig"

今日の目標

- 信頼区間・信用区間・最高事後密度領域の違いと特徴を説明できる
- 事後確率密度から信用区間・最高事後密度領域を求められる



Quiz 解答: ベータ分布

①

$$\begin{aligned} E[\Theta] &= \int_0^1 \theta \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} d\theta \\ &= \frac{a}{a+b} \times \int_0^1 \frac{\Gamma(a+1b)}{\Gamma(a+1)\Gamma(b)} \theta^{(a+1)-1} (1-\theta)^{b-1} d\theta = \frac{a}{a+b}. \end{aligned}$$

② $V[\Theta] = E[\Theta^2] - E[\Theta]^2$. $E[\Theta^2]$ は $E[\Theta]$ と同様に.③ $f'(\theta) = 0$ を θ について解く.

二項モデル（複数のベルヌイ分布）

Hoff ペイズ §3.1

値が不明なパラメタ θ の確率で表 $Y = 1$ が出るコインを N 回投げた結果を, $Y_1, \dots, Y_N = 0, 1$ とする.

Y_i は独立同分布にしたがう. $Y_i \stackrel{\text{i.i.d.}}{\sim} B(1, \theta)$ ($n = 1, p = \theta$ の二項分布).
 θ は不明ということで, 分布をもつ確率変数 Θ と思って, 前回の Quiz と

- $Y_1, \dots, Y_N$ の組 $\leftrightarrow$ 柿の質量 X
- $\Theta \leftrightarrow$ 柿の品種 AB Y

と対応させ, θ の確率分布を $y_1, \dots, y_N$ の情報で更新していきたい.
 $N = 1$ のとき, $Y_1 = y_1$ を得る確率は (なぜか $\text{given } |$ を使って),

$$f_{Y|\theta}(y_1|\theta) = \theta^{y_1}(1 - \theta)^{1-y_1}.$$

一般の N のとき,

$$f_{Y|\theta}(y_1, \dots, y_N|\theta) = \theta^{\sum_i y_i} (1 - \theta)^{N - \sum_i y_i}.$$

θ の分布

θ についての主観的確率（信念）は，連続型確率変数 Θ で表現できる．
何の先入観も持たないときの事前確率密度は $\Theta \sim U(0, 1)$ ，すなわち，一様分布

$$f_{\theta}(\theta) = \begin{cases} 1 & (0 \leq \theta \leq 1) \\ 0 & (\text{他}) \end{cases}$$

と考えられる．

このとき， $f_{Y|\theta}(y_1, \dots, y_N | \theta)$ は条件付き確率であることが納得できる．
ベイズの定理より， $Y_1 = y_1, \dots, Y_N = y_N$ という結果を得たときの Θ の事後確率密度は，

$$\begin{aligned} f_{\theta|Y}(\theta | y_1, \dots, y_N) &= \frac{f(y_1, \dots, y_N, \theta)}{f_Y(y_1, \dots, y_N)} \\ &= \frac{f_{Y|\theta}(y_1, \dots, y_N | \theta) f_{\theta}(\theta)}{f_Y(y_1, \dots, y_N)} \\ &= \frac{\theta^{\sum_i y_i} (1 - \theta)^{N - \sum_i y_i} \cdot 1}{(0 < \theta < 1)} \end{aligned}$$

ここまで来たよ

3 二項モデルとベイズ推定

4 信頼区間・信用区間・最高事後密度領域

- 信用区間・最高事後密度領域の定義
- ポアソンモデル

信頼区間 (confidence interval)

定義 (頻度論的な信頼区間 Hoff ベイズ定義 6,p.47)

上下限が確率変数であるような θ の区間 $[\ell(Y|\theta), u(Y|\theta)]$ が、パラメタ θ に対する、信頼係数 $1 - \alpha$ の信頼区間であるとは、

$$P(\ell(Y|\theta) < \theta < u(Y|\theta)) = 1 - \alpha$$

であることをいう。

これに、実現値 $Y = y$ を代入したものも信頼区間とよばれる。
しばしば、 $P(\theta < \ell(Y)) = P(\theta > u(Y)) = \frac{\alpha}{2}$ であるようにとる。

例

$B(1, \theta)$ のサイズ N の標本からの, θ の区間推定.

標本平均値 $\bar{Y} = \frac{1}{N} \sum_i y_i$. $E[Y] = \theta$, $V[Y] = \theta(1 - \theta)/N$.

$$P(\theta + \Phi^{-1}(\frac{\alpha}{2})\sqrt{\frac{\theta(1 - \theta)}{N}} < \bar{Y} < \theta + \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{\theta(1 - \theta)}{N}}) = 1 - \alpha$$

移項

$$P(\bar{Y} - \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{\theta(1 - \theta)}{N}} < \theta < \bar{Y} - \Phi^{-1}(\frac{\alpha}{2})\sqrt{\frac{\theta(1 - \theta)}{N}}) = 1 - \alpha$$

近似

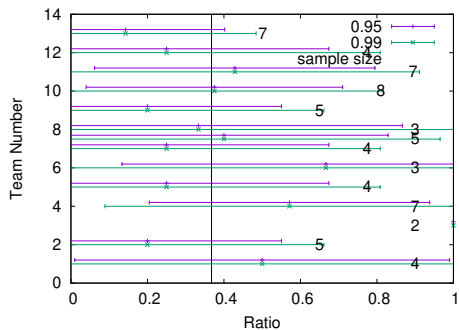
$$P(\bar{Y} - \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{\bar{Y}(1 - \bar{Y})}{N}} < \theta < \bar{Y} - \Phi^{-1}(\frac{\alpha}{2})\sqrt{\frac{\bar{Y}(1 - \bar{Y})}{N}}) = 1 - \alpha$$

L04-Q1

Quiz(母比率の区間推定)

選挙で出口調査をしたところ、50 人中 35 人が A 候補に投票したと答えた。母集団を投票した人 2500 人とする。そのうち A 候補に投票した人の母比率 (得票率) を考える。

- ① A 候補の得票率を, (点) 推定しよう
- ② A 候補の得票率を, 信頼係数 $1 - \alpha = 0.95$ で区間推定しよう。
- ③ A 候補の得票率を, 信頼係数 $1 - \alpha = 0.99$ で区間推定しよう。



信用区間 (credible interval)

定義 (ベイズ信用区間 Hoff ベイズ定義 5,p.47)

観測データ $Y = y$ に基づく θ の区間 $[\ell(y), u(y)]$ が, パラメタ θ に対する, 信頼係数 $1 - \alpha$ の信用区間であるとは,

$$P(\ell(y) < \Theta < u(y)) = 1 - \alpha$$

が成り立つことをいう.

Hoff ベイズ p.48 しばしば, $P(\Theta < \ell(y)) = P(\Theta > u(y)) = \frac{\alpha}{2}$ であるようにとる.

θ_q を $\Theta|Y = y$ の q 分位数とすると, $\ell(y), u(y)$ は, $\frac{\alpha}{2}, 1 - \frac{\alpha}{2}$ 分位数.
figure.

L04-Q2

Quiz(ベータ分布のベイズ信用区間)

$\Theta \sim \text{beta}(a, b)$ のとき、ベイズ推定の結果、パラメタ X の事後確率密度が $\text{beta}(2, 3)$ で与えられた．`scipy.stats.beta` を利用して， X の 0.95 信用区間を数値で求めよう．

最高事後密度 (HPD=highest posterior density) 領域

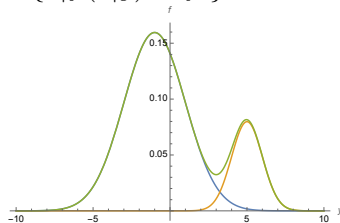
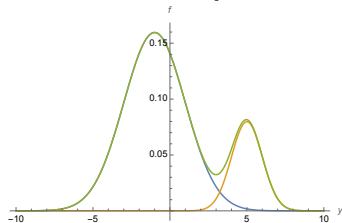
定義 (最高事後密度領域 Hoff バイズ定義 7,p.49)

観測データ $Y = y$ に基づく $(1 - \alpha)$ 最高事後密度領域とは、次を満たすパラメタ空間の部分集合 $s(y)$ のこと.

$$P(\theta \in s(y) | Y = y) = 1 - \alpha$$

$$(\theta_a \in s(y) \text{ かつ } \theta_b \notin s(y)) \Rightarrow f_{\theta|Y}(\theta_a|y) > f_{\theta|Y}(\theta_b|y)$$

ぎりぎりの密度 $f_c \in \mathbb{R}$ で、 $s(y) = \{\theta | f(\theta|y) \geq f_c\}$ の形で書かれるはず.



L04-Q3

Quiz(混合ガウス分布の最高事後密度領域)

ベイズ推定の結果、パラメタ X の事後確率密度が次の混合ガウス分布で与えられた。

$$f(x) = \frac{1}{2} \frac{1}{\sqrt{2\pi} \cdot 2^2} e^{-\frac{(x+6)^2}{2 \cdot 2^2}} + \frac{1}{2} \frac{1}{\sqrt{2\pi} \cdot 2^2} e^{-\frac{(x-6)^2}{2 \cdot 1^2}}.$$

次を、標準正規分布の累積分布関数 $\Phi(z)$ やその逆関数を用いて表そう。

- ① X の 0.95 ベイズ信用区間。
- ② X の 0.95 最高事後密度領域。ただし、ガウス分布の $|x - \mu| > 3\sigma$ の領域においては、確率密度関数の値は 0 であるかのように近似してよい。

L04-Q4

Quiz(指数分布の信用区間と最高事後密度領域)

ベイズ推定の結果，パラメタ X の事後確率密度が次の指数分布で与えられた ($\lambda > 0$).

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & (x \geq 0) \\ 0 & (\text{他}) \end{cases}$$

- ① X の $1 - \alpha = 0.95$ ベイズ信用区間を求めよう.
- ② X の $1 - \alpha = 0.95$ 最高事後密度領域を求めよう.

ここまで来たよ

3 二項モデルとベイズ推定

4 信頼区間・信用区間・最高事後密度領域

- 信用区間・最高事後密度領域の定義
- ポアソンモデル

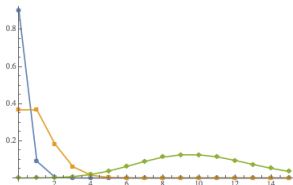
ポアソンモデル

Hoff ベイズ §3.2

ポアソン分布

離散型確率変数 X が次の確率分布を持つとき, X はパラメタ λ のポアソン分布 $\text{Po}(\lambda)$ に従うという.

$$P(X = k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & (k = 0, 1, 2, 3, \dots) \\ 0 & (\text{他}) \end{cases}$$



$\text{Po}(0.1), \text{Po}(1), \text{Po}(10)$

意味: 独立に, 時間に比例して, 単位時間に平均すると λ 回起きる事象が, 単位時間内に k 回起きる確率.

命題 (ポアソン分布の母平均値と母分散)

$$E[X] = \lambda, \quad V[X] = \lambda.$$

L04-Q5

Quiz(ポアソン分布)

ある県では、交通死亡事故が、平均すると1日に3件起きるという。1日の事故の件数はパラメタ $\lambda = 3$ のポアソン分布に従う。

- ① 特定の1日に、交通死亡事故が0件である確率を求めよう。
- ② 特定の1日に、交通死亡事故が6件である確率を求めよう。
- ③ 1日に起きる交通死亡事故の件数の分散を求めよう。

ポアソン分布と二項分布の似ているところ違うところ

	ポアソン分布 $Po(\lambda)$	二項分布 $B(n, p)$
??に k 回		
k の上限		
母平均値		
母分散		

L04-Q6

Quiz(ポアソン分布)

あるサッカーチームは, 1 ゲームで平均 4.5 点得点できる. 1 ゲームで得点が 4, 5, 6 点である確率は?

L04-Q7

Quiz(二項分布)

帝国軍クロントゥルーパーからなる, あるサッカーチームは, PK 成功率が 0.9 である. 5 人が蹴る PK 戦で得点が 4, 5, 6 点である確率は?

L04-Q8

Quiz(ポアソン分布)

あるサッカーチームは、90 分のゲームで平均 3 点得点できる。

- ① ハーフ 45 分間に 0 点である確率は?
- ② ハーフ (前半) 0 点 かつ ハーフ (後半) 3 点である確率は?
- ③ ゲーム 90 分で 3 点であるときに、ハーフ (前半) 0 点, ハーフ (後半) 3 点である確率は?

ポアソンモデルの共役分布は？

$Y_1, \dots, Y_N \stackrel{\text{i.i.d.}}{\sim} \text{Po}(\lambda)$, すなわち,

$$f_{Y|\theta}(y|\theta) = \frac{\theta^y}{y!} e^{-\theta}$$

$$f_{Y|\theta}(y_1, \dots, y_N|\theta) = \frac{\theta^{\sum_i y_i}}{y_1! \dots y_N!} e^{-N\theta}$$

とする.

θ の事前分布が $f_\theta(\theta)$ のとき,

θ の事後分布は

$$f_{\theta|Y}(\theta|y_1, \dots, y_N) = C \times f_{Y|\theta}(y_1, \dots, y_N|\theta) f_\theta(\theta) = \frac{\theta^{\sum_i y_i}}{y_1! \dots y_N!} e^{-N\theta} f_\theta(\theta).$$

共役である（＝事前と事後が同じクラスに属する）のはどんな分布？

$$f(\theta) = C' \times \theta^{a-1} e^{-b\theta}$$

これがガンマ分布.

ガンマ分布

定義 (ガンマ分布)

次の確率密度関数を持つ連続型確率変数 Θ を, パラメタ a, λ のガンマ分布 $\text{gamma}(a, \lambda)$ にしたがうという.

$$f(\theta|a, \lambda) = \begin{cases} \frac{\lambda^a}{\Gamma(a)} \theta^{a-1} e^{-\lambda\theta} & (\theta > 0) \\ 0 & (\text{他}) \end{cases}$$

指数分布は $a = 1$ のガンマ分布.

命題 (ガンマ分布の性質)

$$\begin{aligned} E[\Theta] &= \frac{a}{\lambda} \\ V[\Theta] &= \frac{a}{\lambda^2} \end{aligned}$$