2. 確率とビリーフ(1)

植野真臣
電気通信大学
大学院情報システム学研究科

教科書正誤表

http://www.ai.is.uec.ac.jp/lecture/

1. 確率

定義1 (σ 集合体)
Ωを標本空間(sample space)とし，$\mathcal{A}$が以下の条件を満たすならばσ集合体(σ-field)と呼ぶ.

1. $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$(ただし$A^c = \Omega \setminus A$)
3. $A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

つまり，たがいに素な事象の和集合により新しい事象を生み出すことができ，それらすべての事象を含んだ集合をσ 集合体と呼ぶ.
σ 集合体上で確率(probability)は以下のように定義される.

定義2 (確率測度)
いま，σ 集合体上$\mathcal{A}$で，つぎの条件を満たす測度(measure)Pを，確率測度(probability measure)と呼ぶ(Kolmogorov 1933).

1. $A \in \mathcal{A}$について$0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. たがいに素な事象列$\{A_n\}_{n=1}^{\infty}$に対して，$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$

演習:以下を証明せよ

定理1 (余事象の確率) 事象A の余事象(complementary event)の確率は以下のとおりである.

$$P(A^c) = 1 - P(A)$$

定理2 (境界)

$$P(\phi) = 0$$

定理3 (単調性)

$$A \subset B \text{のとき} P(A) \leq P(B)$$

定理4 (確率の和法則)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

3. 主観確率

ラプラスの頻度論

コインを何百回も投げて表が出た回数(頻度)を数えて，その割合を求めることを考えよう. いま，投げる回数をn とし，表の出た回数n_1とすると，$n \to \infty$のとき，

$$\frac{n_1}{n} \to \frac{1}{2}$$

となることが予想される. このように，何回も実験を繰り返してn 回中，

事象A がn_1 回出たとき，$\frac{n_1}{n}$をA の確率と解釈するのが頻度主義である.

しかし，この定義では真の確率は無限回実験をしなければならないので得ることは不可能である. また，科学的実験が可能な場合にのみ確率が定義され，実際の人間が扱う不確かさに比べてきわめて限定的になってしまう.

4. 主観確率

例えば，松原(2010)では以下のような主観確率の例が挙げられている.

1. 第三次世界大戦が20XX 年までに起こる確率が0.01
2. 明日，会社の株式の価格が上がる確率が0.35
3. 来年の今日，東京で雨が降る確率が0.5

ベイズ統計では，これらの主観確率は個人の意思決定のための信念として定義され，ビリーフ(belief)と呼ばれる. 当然，頻度論的確率を主観確率の一種とみなすことができるが，その逆は成り立たない.

本書では，ベイズ統計の立場に立ち，確率をビリーフの立場で解釈する. ビリーフの具体的な決定の仕方など厳密な理論に興味のある読者はBernardo and Smith (1994), Berger(1985)を参照されたい.

条件付き確率

定義**3** (条件付き確率)
$A \in \mathcal{A}, B \in \mathcal{A}$について，事象$B$ が起こったという条件の下で，事象A が起こる確率を条件付き確率(conditional probablity)と呼び，

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

で示す．このとき，$P(A|B) = \frac{P(A \cap B)}{P(B)}$より以下の乗法公式が成り立つ．

定理**5** (乗法公式)

$P(A \cap B) = P(A|B)P(B)$

このとき，$P(A \cap B)$ をA とB の**同時確率**(joint probability)と呼ぶ．

独立

定義**4** (独立)
ある事象の生起する確率が，他のある事象が生起する確率に依存しないとき，二つの事象は独立(independent)であるという．すなわち事象Aと事象B が独立とは$P(A|B) = P(A)$であり，

$P(A \cap B) = P(A)P(B)$

が成り立つことをいう．

さらに乗法公式を一般化すると以下のチェーンルールが導かれる．

$P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$

これは，3 個以上の事象にも拡張できるので，チェーンルール(chain rule) は以下のように書ける．

定理**6** チェーンルール
N個の事象 $\{A_1, A_2, \cdots, A_N\}$について
$P(A_1 \cap A_2 \cap \cdots \cap A_N)$
$= P(A_1 | A_2 \cap A_3 \cap \cdots \cap A_N)P(A_2 | A_3 \cap A_4 \cap \cdots \cap A_N) \cdots P(A_N)$
が成り立つ．

5．全確率の定理

定理**7** (全確率の定理(total probability theorem))

たがいに背反な事象$A_1, A_2, \cdots, A_n$ $(A_i \in \mathcal{A})$ が全事象Ωを分割しているとき，
事象$B \in \mathcal{A}$について，
$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$ が成り立つ．

6．ベイズの定理

定理**8** (ベイズの定理(Bayes' theorem))
たがいに背反な事象$A_1, A_2, \cdots, A_n$ が全事象Ωを分割しているとする．
このとき，事象$B \in \mathcal{A}$について，

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

が成り立つ．

例題

例**1** 昔，ある村にうそつき少年がいた．少年はいつも「オオカミが来た！！」と大声で叫んでいたが，いままで本当だったことがない．

「オオカミが来た」という事象をA
少年が「オオカミが来た！！」と叫ぶ事象をB

とし，$P(B|A) = 1.0$, $P(B|A^c) = 0.5$, $P(A) = 0.005$ とする．少年が「オオカミが来た！！」と叫んだとき実際にオオカミが来ている確率を求めてみよう．

7．ビリーフ

ベイズ統計では，より広い確率の解釈として「ビリーフ」(belief)を用いることは先に述べた．ここでは考え方のみについてふれよう．意思決定問題から個人的な主観確率であるビリーフが以下のように求められる．
例えば，つぎの二つの賭けを考えよう．

1. もしオオカミが来ていれば1 万円もらえる
2. 赤玉n 個，白玉$100-n$ 個が入っている合計100 個の玉が入っている壺の中から一つ玉を抜き出し，それが赤玉なら1 万円もらえる．

どちらの賭けを選ぶかといわれれば，2 番目の賭けで赤玉が100 個ならば，誰もが迷わず2 番目の賭けを選ぶだろうし，逆に$n = 0$ ならば，1 番目の賭けを選ぶだろう．この二つの賭けがちょうど同等になるようにn を設定することができれば，$\frac{n}{100}$があなたの「オオカミが来る」ビリーフになる．このように，ベイズ統計における確率の解釈「ビリーフ」は頻度主義の確率で扱える対象を拡張でき，個人的な信念やそれに基づく意思決定をも合理的に扱えるツールとなる．

ビリーフを用いてもう一度例を振り返ろう．例1．では，もともとのオオカミが来る確率$P(A) = 0.005$が，（うそかどうかわからない）少年の報告により
$P(A \mid B) = 0.00995$と約2 倍にビリーフが更新されていることがわかる．すなわち，うそをつく少年の証言によって事前のビリーフが事後のビリーフに更新されたのである．
このとき，ベイズ統計では，

少年の証言を「エビデンス」（evidence）と呼び，
事前のビリーフを「**事前確率**」（prior probability），
事後のビリーフを「**事後確率**」（posterior probability）

と呼ぶ

例題

例**2** (**3** 囚人問題)
つぎに有名な3 囚人問題を紹介しよう．ある監獄にアラン，バーナード，チャールズという3 人の囚人がいて，それぞれ独房に入れられている．3 人は近く処刑される予定になっていたが，恩赦が出て3 人のうち1人だけ釈放されることになったという．

誰が恩赦になるかは明かされておらず，それぞれの囚人が「私は釈放されるのか?」と聞いても看守は答えない．

囚人アランは一計を案じ，看守に向かって「私以外の2 人のうち少なくとも1 人は死刑になるはずだ．その者の名前が知りたい．私のことじゃないんだから教えてくれてもよいだろう?」と頼んだ．

すると看守は「バーナードは死刑になる」と教えてくれた．それを聞いたアランは「これで釈放される確率が1/3 から1/2に上がった」とひそかに喜んだ．果たしてアランが喜んだのは正しいのか?

8．確率変数

一つの試行の結果を標本点$\omega \in \Omega$と呼ぶ．
この標本点$\omega \in \Omega$は，なんらかの測定によって観測される．
この測定のことを，確率論では**確率変数**(randomvariable)と呼ぶ．

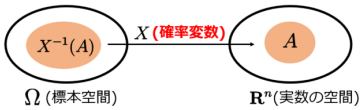
例えば，コインをn回投げるという試行について，表が出る回数Xは確率変数である．このとき，標本点ωは表・裏のパターンがn個あり得るので2^n通りあり，Xは0 からnまでの値をとる．
数学的には，確率変数は以下のように定義される．

定義**5** 確率空間(Ω, A, P)に対し，Ω（取り得るすべての値）から実数R（取り出された部分）への関数$X: \Omega \to R$が，任意の実数rに対し
$\{X \leq r\} \in A$（累積値が有限）を満たすならば，

Xを確率空間(Ω, A, P)上の確率変数という．

確率変数

確率変数は標本空間から実数空間への関数（写像）。



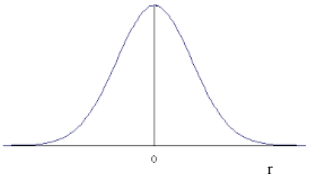
$X^{-1}(A) := \{\omega \in \Omega \mid X(\omega) \in A\}$

写像 $X : \Omega \to \mathbf{R}^n$ が**確率変数**であるとは
任意の$A \in \mathcal{B}(\mathbf{R}^n)$ に対して $X^{-1}(A) \in \mathcal{F}$ を満たすときにいう。

ボレル集合族 (σ加法族)　　σ 加法族

例

工場でボルトを生成している。
ボルトの長さのあり得る値をΩ
を実数空間に写像する関数が確率変数



9．分布関数

確率変数Xが確率空間(Ω, A, P)上に定義されると，任意の$r \in R$に対して

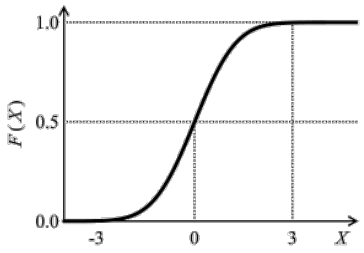
$F(r) = P(\{X \leq r\})$

が定まる．この$F(r)$はRからRへの関数であり，**分布関数**（distribution function）と呼ばれる．分布関数$F(r)$の性質は，

1．rに対して単調増加，
2．$r \to -\infty$と$r \to \infty$のとき，それぞれ，極限値0 と1 をもつ，
3．$F(r)$は右半連続，すなわち，$h \to 0$のとき$F(r+h) - F(r) \to 0$

であり，逆にこれらの性質を満たせば，それを分布関数としてもつ確率変数が存在することが知られている．

分布関数の例



10. 確率分布

定義**6** 確率変数Xが，たかだか加算個の実数の集合$\chi = \{x_1, x_2, \cdots\}$の中の値をとるならば，$X$は離散であるという．すなわち，Rが加算のとき．

離散確率変数Xのとり得る値$x_k \in \chi$を，その確率$p = P(X = x_k)$に対応づける写像$p: \chi \to [0, 1]$をXの離散確率分布（discrete probability distribution）と呼ぶ．

離散確率分布pは

$p(x) \geq 0 \ (x \in \chi)$ ，$\sum_{x \in \chi} p(x) = 1$

を満たす．逆に，これら二つの条件を満たすX上の関数pを確率分布としてもつ確率変数が存在する．

11. 確率密度関数

定義**7** 確率変数Xが，実数全体の集合$\chi = R$の中の値をとるならば，Xは連続であるという．

$f(x) \geq 0 \ (x \in R)$ ，$\int_{-\infty}^{\infty} f(x) = 1$

を満たし，かつ任意の$a < b$に対して

$p(a < X < b) = \int_a^b f(x)dx$

であるとき，$f(x)$は連続確率変数Xの確率密度関数（probability density function）であるという．

12. 同時確率分布

定義**8**

いま，m個の確率変数をもつ確率分布$p(x_1, x_2, \cdots, x_m)$を変数$x_1, x_2, \cdots, x_m$ **の同時確率分布**（joint probability distribution）と呼ぶ．

13. 周辺確率分布

定義**9**

x_iのみに興味がある場合，同時確率分布からx_iの確率分布は，離散型の場合，

$p(x_i) = \sum_{x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_m} p(x_1, x_2, \cdots, x_m)$

連続型の場合，

$p(x_i) = \int p(x_1, x_2, \cdots, x_m) dx_1, \cdots, dx_{i-1}, dx_{i+1}, \cdots, dx_m$

で求められ，$p(xi)$を離散型の場合，**周辺確率分布**（marginal probability distribution），連続型の場合，**周辺密度関数**（marginal probability density function）と呼ぶ．

14. 確率分布とパラメータ

定義**10** (パラメータ空間と確率分布)
k次元パラメータ集合を$\boldsymbol{\Theta} = \{\theta_1, \theta_2, \cdots, \theta_k\}$と書くとき，確率分布は以下のような関数で示される．

$$f(x|\boldsymbol{\Theta})$$

すなわち，確率分布$f(x|\boldsymbol{\Theta})$の形状はパラメータ$\boldsymbol{\Theta}$のみによって決定され，パラメータΘのみが確率分布$f(x|\boldsymbol{\Theta})$を決定する情報である．

例**3** コインをn回投げたとき，表が出る回数を確率変数xとした確率分布は以下の二項分布に従う．

$$f(x|\theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

ここで，θは，コインの表が出る確率のパラメータを示す．

15. 尤度原理

定義**11** (尤度) $X = (X_1, \cdots, X_i, \cdots, X_n)$ が確率分布 $f(X_i|\theta)$ に従うn個の確率変数とする.
n 個の確率変数に対応したデータ $\boldsymbol{x} = (x_1, \cdots, x_n)$ が得られたとき,

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

を尤度関数(likelihood function)と定義する(Fisher, 1925).

尤度の例

例**4** コインをn 回投げたとき, 表が出た回数がx 回であったときのコインの表が出るパラメータθ の尤度は

$L(\theta|n,x) \propto \binom{n}{x} \theta^x (1-\theta)^{n-x}$

もしくは,

$L(\theta|n,x) \propto \binom{n}{x} \theta^x (1-\theta)^{n-x}$

でもよい.

尤度は, データパターンが観測される確率に比例する, パラメータθ の関数である
尤度は確率の定義を満たす保証がないために確率とは呼べないが, これを厳密に確率分布として扱うアプローチが後述するベイズアプローチである

尤度を最大にするパラメータθ を求めることは, データを生じさせる確率を最大にするパラメータθ を求めることになり, その方法を**最尤推定法**(maximum likelihood estimation, **MLE**)と呼ぶ.

最尤推定値

定義**12** (最尤推定量)
データxを所与として, 以下の尤度最大となるパラメータを求めるとき,

$$L(\theta|x) = \max\{L(\theta|x) : \boldsymbol{\theta} \in C\}$$

$\hat{\theta}$を最尤推定量(maximum likelihood estimator)と呼ぶ(Fisher 1925).

ただし, C はコンパクト集合を示す.

例題

例**5** (二項分布の最尤推定)

コインを投げてn 回中x 回表が出たときの確率θ の最尤推定値を求めよう.

例題

例**6** (正規分布)

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\}$$

について, データ$(x_1, \cdots, x_n)$を得たときの平均値パラメータμ, および分散パラメータσ^2 の最尤推定値を求めよう.

強一致性

定義**13** (強一致性)
推定値$\hat{\theta}$が真のパラメータθ^*に概収束するとき, $\hat{\theta}$は強一致推定値(strongly consistent estimator)であるという.

$$P(\lim_{n\to\infty} \hat{\theta} = \theta^*) = 1.0$$

つまり, データ数が大きくなると推定値が必ず真の値に近づいていくとき, その推定量を強一致推定値と呼ぶ.

最尤推定値の一致性

定理**9** (最尤推定値の一致性)
最尤推定値$\hat{\theta}$は真のパラメータθ^*の強一致推定値である（Wald,1949）.

最尤推定値の漸近正規性

定義**14**
θ^*の推定値 $\hat{\theta}$が**漸近正規推定量**（asymptotically normal estimator）であるとは，$\sqrt{n}(\hat{\theta}-\theta^*)$ の分布が正規分布に分布収束することをいう. すなわち，任意の$\theta^* \in \Theta^*$ と任意の実数に対して

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(\hat{\theta}-\theta^*)}{\sigma(\theta^*)} \leq x\right) = \Phi(x)$$

このことを，$\sqrt{n}(\hat{\theta}-\theta^*) \xrightarrow{as} N(0, \sigma^2(\theta^*))$と書く.
$\sigma^2(\theta^*)$ を漸近分散（asymptotic variance）という.

最尤推定値の漸近正規性

定理**10**
確率密度関数が正則条件（regular condition）の下で，微分可能のとき，
最尤推定量は漸近分散$I(\theta^*)^{-1}$をもつ漸近正規推定量である.

$$I(\theta^*) = E_\theta\left[\left(\frac{\partial}{\partial\theta} lnL(\theta|\mathbf{x})\right)^2\right]$$

をフィッシャー（Fischer）の情報量と呼ぶ.