『社会科学のためのベイズ統計モデリング』第 5 章

Hiroshi Hamada
Tohoku University

May, 2020
at Tohoku University

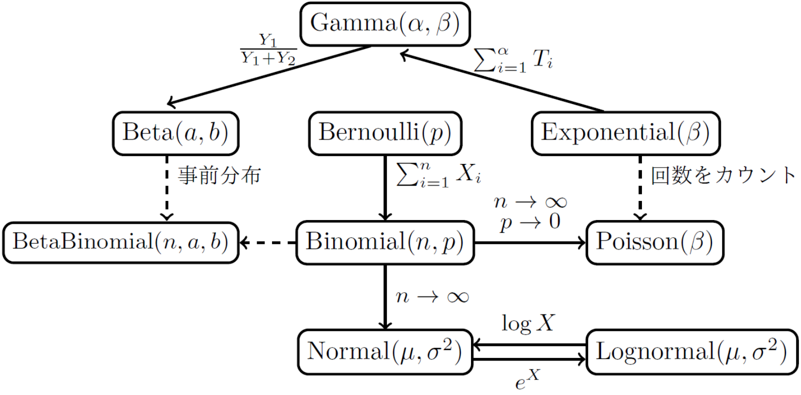
Gamma(α, β)
$\frac{Y_1}{Y_1+Y_2}$
$\sum_{i=1}^{\alpha} T_i$
Beta(a, b)
Bernoulli(p)
Exponential(β)
事前分布
$\sum_{i=1}^{n} X_i$
回数をカウント
$n \to \infty$
$p \to 0$
BetaBinomial(n, a, b)
Binomial(n, p)
Poisson(β)
$n \to \infty$
$\log X$
Normal(μ, σ^2)
Lognormal(μ, σ^2)
e^X

図 5.1　確率分布の関係

ベルヌーイ分布

- 実現値: $y \in \{0, 1\}$
- パラメータ: $q \in [0, 1]$
- 確率質量関数: $\mathrm{Bernoulli}(y|q) = q^y(1-q)^{1-y}$
- 平均: q，標準偏差: $\sqrt{q(1-q)}$

注目する事象が確率 q で起こったとき 1，確率 $1-q$ で起こらなかったとき 0 の値をとる確率変数の分布を**ベルヌーイ分布**（Bernoulli distribution）という．

ベルヌーイ試行

- 試行の結果は成功か失敗のいずれかである
- 各試行は独立である
- 成功確率 q，失敗確率 $1-q$ は試行を通じて一定である

2 項分布

- 実現値: $x \in \{0, 1, 2, \ldots, n\}$
- パラメータ: $n \in \mathbb{Z}^+$, $p \in [0, 1]$
- 確率質量関数: $\mathrm{Binomial}(x|n, p) = {}_nC_x p^x (1-p)^{n-x}$
- 平均: np，標準偏差: $\sqrt{np(1-p)}$

注目する事象が確率 q で生じるベルヌーイ試行を n 回繰り返したとき，その事象が起こった回数 x は 2 項分布（binomial distribution）に従う

例）コインを投げて n 回中 x 回表がでる確率は 2 項分布で決まる.

命題 (確率変数のたたみこみ)

X, Y を独立な確率変数とし，$Z = X + Y$ とおく．
X, Y, Z の確率密度（質量）関数を $f(x), g(y), h(z)$ とおく．

- 離散確率変数の場合 $h(z) = \sum_x f(x)g(z-x)$
- 連続確率変数の場合 $h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx$

h を f と g のたたみこみといい，$h(z) = f * g(z)$ と表す．ただし X, Y が非負の値をとる場合の和と積分の範囲は $x = 0$ から $x = z$ までとする．

ポアソン分布

- 実現値: $x \in \{0, 1, 2, \ldots\}$
- パラメータ: $\lambda \in \mathbb{R}^+$
- 確率質量関数: $\mathrm{Poisson}(x|\lambda) = \dfrac{\lambda^x}{x!} e^{-\lambda}$
- 平均: λ，標準偏差: $\sqrt{\lambda}$

ポアソン分布（Poisson distribution）は，単位時間当たりに注目する事象が生じる回数の確率分布として，よく使われる．

例）ウェブ上に公開されたブログへの 1 日あたりのアクセス人数の分布

ポアソン分布の 2 種類の導出

命題

2 項分布 $\mathrm{Binomial}(n, p)$ は $np = \lambda$ を一定に保って n を限りなく大きくすると，ポアソン分布で近似できる．

命題

ポアソン過程から，単位時間内にイベントが生じる回数の分布としてポアソン分布を導出できる．

指数分布

- 実現値: $x \geq 0$ を満たす実数 x
- パラメータ: $\lambda \in \mathbb{R}^+$
- 確率密度関数: $\mathrm{Exponential}(x|\lambda) = \lambda e^{-\lambda x}$
- 平均: $1/\lambda$，標準偏差: $1/\lambda$

指数分布（exponential distribution）は，注目する事象が特定の条件下で起きるまでの時間の分布を表す．

例）災害が起こった直後から次の災害が起こるまでの時間や，商品を使い始めてから壊れるまでの時間．

無記憶性

確率変数 X が任意の $s > 0, t > 0$ について

$$P(X > s + t | X > t) = P(X > s)$$

を満たすことを無記憶性（memorylessness）という．
指数分布は「無記憶性」を満たす確率変数が従う分布である．

ポアソン分布と指数分布

ポアソン分布 時間区間 $(0, t]$ 内に事象が生じる回数の分布

指数分布 ポアソン分布にしたがう事象が 1 回発生するまでの時間の分布

正規分布

- 実現値: $x \in \mathbb{R}$
- パラメータ: $\mu \in \mathbb{R},\ \sigma \in \mathbb{R}^+$
- 確率密度関数:
$$\mathrm{Normal}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$
- 平均: μ，標準偏差: σ

命題 (ド・モアブル–ラプラスの中心極限定理)

パラメータ p のベルヌーイ分布に従う確率変数 $X_1, X_2, \ldots, X_n$ が互いに独立であるとし,

$$S_n = \frac{(X_1 + X_2 + \cdots + X_n) - np}{\sqrt{np(1-p)}}$$

とおく．確率変数 S_n は $n \to \infty$ のとき，平均 0 で標準偏差 1 の正規分布に従う

命題の意味

$$S_n = \frac{(X_1 + X_2 + \cdots + X_n) - np}{\sqrt{np(1-p)}}$$

- S_n の分子にある $X_1 + X_2 + \cdots + X_n$ の部分は，ベルヌーイ分布を n 個足しあわせた確率変数で 2 項分布に従う.
- 2 項分布の平均と標準偏差は np, $\sqrt{np(1-p)}$ である.
- S_n は 2 項分布に従う確率変数を，その平均 np と標準偏差 $\sqrt{np(1-p)}$ で標準化した確率変数である.

より一般的な中心極限定理

命題 (中心極限定理)

平均 μ，σ^2 であるような独立同分布に従う確率変数

$$X_1, X_2, \ldots, X_n$$

を考え，ある $0 < \delta < 1$ が存在して任意の i について $\mathbb{E}[|X_i - \mu|^{2+\delta}] = K < +\infty$ が成立すると仮定する．このとき確率変数

$$\frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}}$$

は $n \to \infty$ のとき，平均 0 で標準偏差 1 の正規分布（標準正規分布）に従う

- 確率的に変動する量 $X_1, X_2, \ldots, X_n$ を合計した X の分散に比して，各 X_j の分散が十分に小さければ，X の分布は正規分布で近似できる
- 正規分布は『適度な大きさの分散をもつ確率変数をたくさん足し合わせて基準化した確率変数が従う分布』
- 背後に中心極限定理の成立が想定できる場合は、正規分布を仮定する一応の根拠となる

対数正規分布

- 実現値: $y \in \mathbb{R}^+$
- パラメータ: $\mu \in \mathbb{R},\ \ \sigma \in \mathbb{R}^+$
- 確率密度関数:

$$\text{Lognormal}(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma y}\exp\left\{-\frac{(\log y-\mu)^2}{2\sigma^2}\right\}$$

- 平均: $\exp\{\mu+\frac{\sigma^2}{2}\}$，標準偏差: $\exp\{\mu+\frac{\sigma^2}{2}\}\sqrt{e^{\sigma^2}-1}$

確率変数 X が平均 μ，標準偏差 σ の正規分布に従っているとき，$Y=e^X$ と定義すると，確率変数 Y の分布は対数正規分布（lognormal distribution）に従う．

置換積分（重要）

$$
\begin{aligned}
P(Y < a) &= \int_0^a \frac{1}{\sqrt{2\pi}\sigma y} \exp\left\{-\frac{(\log y - \mu)^2}{2\sigma^2}\right\} dy \\
&= \int_{\log(0)}^{\log(a)} \frac{1}{\sqrt{2\pi}\sigma e^x} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} e^x dx \\
&= \int_{-\infty}^{\log(a)} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} dx \\
&= P(X < \log a)
\end{aligned}
$$

$X = \log Y$ と変換すれば，X は正規分布に従う

対数正規分布の導出

- アタリが出ると所持金が e 倍になるギャンブル
- 最初の所持金を 1 円とすれば，x 回アタリがでたときの所持金は e^x．各回の試行でアタリが出る確率を p とおけば，n 回中 x 回アタリがでる確率は 2 項分布 $\mathrm{Binomial}(n, p)$ に従う
- n が十分に大きいとき X の分布は正規分布に近づく
- アタリ回数 X が正規分布に従うと仮定すると，所持金 Y は確率変数 $Y = e^X$．ゆえに所持金の分布は, 対数正規分布で表せる.

$$Y \sim \mathrm{Lognormal}(np, \sqrt{np(1-p)})$$

ベータ分布

- 実現値: $x \in (0, 1)$
- パラメータ: $\quad a \in \mathbb{R}^+, b \in \mathbb{R}^+$
- 確率密度関数: $\mathrm{Beta}(x|a, b) = \dfrac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a, b)}$
- 平均: $\dfrac{a}{a+b}$, 標準偏差: $\dfrac{\sqrt{ab}}{(a+b)\sqrt{a+b+1}}$

ベータ分布（Beta distribution）は実現値が区間 $(0, 1)$ に収まるような連続確率変数の分布

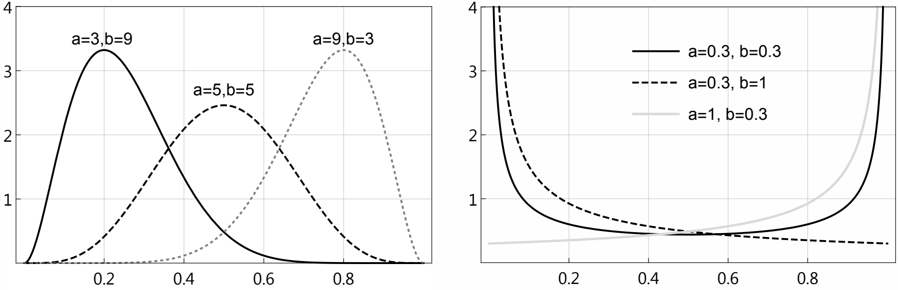

Figure: ベータ分布の確率密度関数

パラメータ次第でさまざまな形状に変化する．a, b 単体で平均や標準偏差の意味はない．確率分布のパラメータは常にモーメント（の関数）に一致するわけではない．

ベータ 2 項分布

- 実現値: $x \in \{0, 1, 2, \dots, n\}$
- パラメータ: $a \in \mathbb{R}^+, b \in \mathbb{R}^+, n \in \mathbb{Z}^+$
- 確率密度関数:

$$\mathrm{BetaBinomial}(x|a, b, n) = {}_nC_x \frac{\mathrm{B}(a + x, b + n - x)}{\mathrm{B}(a, b)}$$

- 平均: $\dfrac{an}{a + b}$, 標準偏差: $\dfrac{\sqrt{abn(a + b + n)}}{(a + b)\sqrt{a + b + 1}}$

ベータ 2 項分布（Beta binomial distribution）は，ベータ分布と 2 項分布を組み合わせた分布.

モデリング

X がパラメータ n, p を持つ 2 項分布にしたがい，さらに p がパラメータ a, b を持つベータ分布に従うと仮定する．

$$
\begin{aligned}
X &\sim \mathrm{Binomial}(n, p) \\
p &\sim \mathrm{Beta}(a, b)
\end{aligned}
$$

分布同士を組み合わせて新たな分布をつくる操作は，統計モデリングでは役立つ！

確率質量関数の導出

$$
\begin{aligned}
f(x) &= \int_0^1 f(x,p)dp \\
&= \int_0^1 {}_nC_x p^x (1-p)^{n-x} \cdot \frac{1}{\mathrm{B}(a,b)} p^{a-1} (1-p)^{b-1}\, dp \\
&= \frac{1}{\mathrm{B}(a,b)} {}_nC_x \int_0^1 p^x (1-p)^{n-x} p^{a-1} (1-p)^{b-1}\, dp \\
&= \frac{1}{\mathrm{B}(a,b)} {}_nC_x \int_0^1 p^{x+a-1} (1-p)^{n-x+b-1}\, dp \\
&= {}_nC_x \frac{\mathrm{B}(a+x, b+n-x)}{\mathrm{B}(a,b)}.
\end{aligned}
$$

ガンマ分布

- 実現値: $y \in \mathbb{R}^+$
- パラメータ: $\alpha \in \mathbb{R}^+,\ \beta \in \mathbb{R}^+$
- 確率密度関数: $\mathrm{Gamma}(x|\alpha, \beta) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$
- 平均: α/β，標準偏差: $\sqrt{\alpha}/\beta$

ガンマ分布（Gamma distribution）は指数分布（ある事象が発生するまでの時間の分布）の和の分布

あるイベントが生じるまでの時間 T が，パラメータ β の指数分布にしたがう．

$$T \sim \mathrm{Exponential}(\beta)$$

確率変数 T が α 個あって，互いに独立であると仮定する．$T_1, T_2, \ldots, T_\alpha$ の和で新しい確率変数 Y をつくる．

$$Y = T_1 + T_2 + \cdots + T_\alpha$$

Y はパラメータ α, β のガンマ分布に従う ($Y \sim \mathrm{Gamma}(\alpha, \beta)$)．

たたみこみ定理

$$Y = T_1 + T_2.$$

T_1, T_2, Y の確率密度関数を $f(t_1), g(t_2), h(y)$ とおく

$$\begin{aligned} h(y) &= \int_0^y f(t_1)g(y-t_1)dt_1 = \int_0^y \lambda e^{-\lambda t_1}\lambda e^{-\lambda(y-t_1)}dt_1 \\ &= \lambda^2 \int_0^y e^{-\lambda t_1}e^{-\lambda(y-t_1)}dt_1 = \lambda^2 \int_0^y e^{-\lambda t_1-\lambda(y-t_1)}dt_1 \\ &= \lambda^2 \int_0^y e^{-\lambda y}dt_1 = \lambda^2 e^{-\lambda y}\int_0^y 1\, dt_1 \\ &= \lambda^2 e^{-\lambda y}\left[t_1\right]_0^y = \lambda^2 e^{-\lambda y} y \end{aligned}$$

確率変数の和で新しい分布を作る→たたみこみ定理（重要）

5 章まとめ

- ベルヌーイ分布: あらゆる現象の基礎．汎用度高し
- 2 項分布: 使いやすい．まずはここからモデリング
- ポアソン分布: レアあるいはランダムな事象の回数に
- 指数分布: 何かが起こるまでの時間．
- 正規分布: 中心極限定理が適用できる場合に．定番
- ベータ分布: パラメータで変幻自在．確率も表現可能
- 対数正規分布: 指数的に増加する量の表現に
- ベータ 2 項分布: 2 項分布を拡張したい時に
- ガンマ分布: 指数分布の合成に．ポアソン分布の共役事前分布として