

多次元確率分布

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 L01(2024-09-25 Wed)

最終更新: Time-stamp: "2024-09-25 Wed 07:49 JST hig"

今日の目標

- 多次元の連続型確率分布の確率密度関数, 累積分布関数, 分位数を説明できる
- 離散-連続混合型の多次元確率分布を説明できる



ここまで来たよ

● はじめに

- この授業どんなのり?

1 多次元確率分布

- 1 変数の連続型確率変数
- 2 次元の確率分布
- 周辺分布と独立性
- 離散-連続混合型

学習目標など

到達目標 → シラバス

- I. ベイズ統計統計ソフトウェアを用いてベイズ統計学に基づいて標本からベイズ的な推定を行える
- II. マルコフ連鎖モンテカルロ法マルコフ連鎖モンテカルロ法を適用して確率変数の期待値を近似的に求められる

I ベイズ推定, と思っていいです.

II は I(や他の目的) のための計算手段の話. ただし, マルコフ連鎖は, モンテカルロ法やベイズ統計に限らず, 必然的に確率論・確率過程的に興味深い対象です.

資料など

<https://hig3.net>

- → 計算科学特論 (配布資料).
- → LearnMoodle <https://learn.hig3.net>
(Google でログイン) → 計算科学特論
- ▶ → Teams



<https://hig3.net>

参考書

- I, II. Hoff ベイズ ピーター D ホフ 標準ベイズ統計学 朝倉書店 2022
- II. 鎌谷 MC 鎌谷研吾 モンテカルロ統計計算 講談社 2020

授業のやり方とピーナッツカウント

授業のり

1-534 で対面，スクリーンと Teams 画面共有下で，PDF 資料に GoodNotes で書き込み，録画公開．GoodNotes の共有 URL は Web から閲覧，アプリから編集可能．

資料はたぶん直前まで完成しません．印刷しません．ノート PC 持参お願いします．

ピーナッツ計算のり

授業内または次の授業までに Moodle にアップロードする（主に手計算）課題 12.5x8.

欠席届

毎回出席を前提に進めます．やむを得ず欠席する場合や，課題を提出できない場合で，ピーナッツ的に考慮されたい場合は事後 2 週間以内に LearnMoodle にある専用のリンクから届けてください．

担当者ののり

- なまえ: 樋口さぶろお hig@math.ryukoku.ac.jp メールはいつでも．
- へや: 1-507
- Web ページ: <https://hig3.net>
- オフィスアワー 後期月昼, 1-539 or Teams chat a00010

ここまで来たよ



はじめに

- この授業どんなのり？



1 多次元確率分布

- 1 変数の連続型確率変数
- 2 次元の確率分布
- 周辺分布と独立性
- 離散-連続混合型

確率密度関数と累積分布関数

岩薩林 確率・統計 §4.1

Hoff ベイズ p.22

X : 実数に値をとる確率変数 (確率空間から実数への関数)
確率 $P(X \text{ の条件})$ が考えられる (大部分の条件に対して)

定義 (確率密度関数 (pdf=probability density function))

確率密度関数 $f: \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ があって, 次のように書ける.

$$P(a < X \leq b) = \int_a^b f(x) \, dx.$$

定義 (母期待値)

X : 連続型確率変数, $g(x)$: 関数に対して, $g(X)$ の母期待値

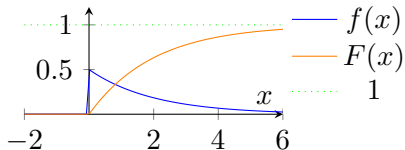
$$E[g(X)] := \int_{-\infty}^{+\infty} g(x) f(x) dx.$$

累積分布関数

定義 (累積分布関数 (cdf=cumulative density function))

$$F(x) := \int_{-\infty}^x f(x') \, dx'.$$

指数分布の確率密度関数と累積分布関数



命題 (累積分布関数の性質)

- $F(x)$ は広義単調増加.
- $P(X \leq b) = F(b)$.
- $P(a < X \leq b) = F(b) - F(a)$
- $F(-\infty) = 0, \quad F(+\infty) = 1$.

経験累積分布関数

定義 (経験累積分布関数 (empirical cumulative density function))

X の標本 $\{x_i\}_{i=1,\dots,N}$ に対して,

$$\hat{F}_N(x) := \frac{|\{i|x_i \leq x\}|}{N}$$

確率密度関数/ヒストグラムのほうが直観的だが (?), 累積分布関数のほうが数学的に扱いやすい

母ナントカ	標本ナントカ
確率密度関数	ヒストグラム
累積分布関数	‘標本累積分布関数’=経験累積分布関数

ヒストグラムは階級によるし, $N \rightarrow +\infty$ で母に似ない.
経験累積は $N \rightarrow +\infty$ で, ある意味, 母に収束する.

分位数

四分位数 (quartile), 中央値 (median) の一般化.

定義 (分位数 (quantile, percentile))

F が単射とする. 連続型確率変数の p 分位数 ($0 \leq p \leq 1$) とは,

$$x = F^{-1}(p).$$

すなわち, 与えられた p に対して,

$$P(X \leq x) = p$$

となる x のこと.

F が単射でないときも, $x = \inf\{x | P(X \leq x) = p\}$ などと定めればあいまいさはない. があまり気にかけてはいない.

Python の scipy.stats での確率変数の扱い

loc(ation)=分布の位置, scale=分布の幅

指数分布の確率分布名は `expon(loc=0,scale=1/lambda)`

```
1  from scipy import stats
2  # 確率密度関数(連続型用) probability density function
3  stats.確率分布名(loc=位置, scale=幅).pdf(x)
4  # 確率(質量)関数(離散型用) probability mass function
5  stats.確率分布名(loc=位置, scale=幅).pmf(x)
6  # 累積分布関数 cummulative distribution function
7  stats.確率分布名(loc=位置, scale=幅).cdf(x)
8  # 分位数=累積分布関数の逆関数 percent point function
9  stats.確率分布名(loc=位置, scale=幅).ppf(x)
10 # サンプル random variable sample
11 stats.確率分布名(loc=位置, scale=幅).rvs(size=サンプルサイズ)
12 確率分布にしたがうランダムな数を サンプルサイズ個 取得する.
13 # 母平均値 mean
14 stats.確率分布名(loc=位置, scale=幅).mean()
15 # 母分散 variance
16 stats.確率分布名(loc=位置, scale=幅).var()
17 # 母標準偏差 standard deviation
18 stats.確率分布名(loc=位置, scale=幅).std()
19 # モーメント  $x$  の  $k$  乗の母期待値
20 stats.確率分布名(loc=位置, scale=幅).moment(k)
21 #  $x$  の一般の関数の母期待値
22 stats.確率分布名(loc=位置, scale=幅).expect(lambdax: f(x))
```

sympy.stats パッケージ, 名前空間

(連続型) 確率分布オブジェクト=`uniform, expon, norm, ...`

(離散型) 確率分布オブジェクト=`binom, poisson, ...`

`.pdf, .cdf` メンバー関数

例: 指数分布

岩薩林 確率・統計 p.80

例 (指数分布)

連続型確率変数 X がパラメタ $\lambda > 0$ に対して確率密度関数

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

を持つとき, X はパラメタ λ の指数分布 (パラメタ $n = 1, \lambda$ のガンマ分布) にしたがうという (分布の幅のパラメタ $a = 1/\lambda$ で指定することもある).

命題 (指数分布のモーメント)

$$E[X^k] = k! \lambda^{-k}, \quad V[X] = \lambda^{-2}.$$

L01-Q1

Quiz(指数分布の確率)

連続型確率変数 X はパラメタ $\lambda = 2$ の指数分布にしたがう.

- ① 確率 $P(X \leq 3)$ を小数で求めよう.
- ② 累積分布関数 $F(x)$ を求めよう.
- ③ 累積分布関数 $F(x)$ の逆関数 $x = F^{-1}(p)$ を求めよう.
- ④ 確率 $P(3 < X \leq 4)$ を小数で求めよう.
- ⑤ 確率 $P(X \geq 4)$ を小数で求めよう.
- ⑥ 確率 $P(X \leq x) = 3/4$ となる x を小数で求めよう.

ここまで来たよ



はじめに

- この授業どんなのり？



1 多次元確率分布

- 1変数の連続型確率変数
- 2次元の確率分布
- 周辺分布と独立性
- 離散-連続混合型

2次元の連続型確率変数

岩薩林 確率・統計 §4.6

Hoff ベイズ p.27

X, Y : 連続型確率変数, $D \subset \mathbb{R}^2$ の, ある性質を満たす部分集合.

定義 (同時確率密度関数)

次の性質を持つ関数 $f: \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}$ を同時確率密度関数という.

$$P((X, Y) \in D) = \iint_D f(x, y) dx dy.$$

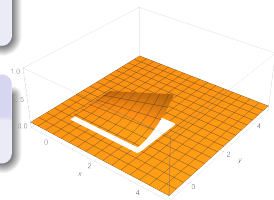
定義 (同時累積分布関数)

$$F(x, y) := P(X \leq x, Y \leq y).$$

便利な分位数は定義できない...

命題 (2次元分布での確率と累積分布関数との関係)

$$P(a < X \leq b, c < Y \leq d) = F(b, d) - F(a, d) - F(b, c) + F(a, c)$$



定義 (母期待値)

X, Y : 連続型確率変数, $g(x, y)$: 関数に対して, $g(X, Y)$ の母期待値

$$E[g(X, Y)] := \iint_{\mathbb{R}^2} g(x, y) f(x, y) dx dy.$$

L01-Q2

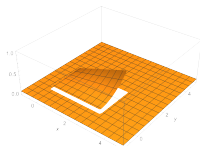
Quiz(独立と限らない多次元の連続型確率変数)

連続型確率変数 X, Y の同時確率密度関数は

$$f(x, y) = \begin{cases} \frac{1}{24}(2 + xy^2) & (0 \leq x \leq 3, 0 \leq y \leq 2) \\ 0 & (\text{他}) \end{cases}$$

で与えられる.

- ❶ 母期待値 $E[X^3Y]$ を求めよう.
- ❷ 確率 $P(X \leq 2 \text{ かつ } Y \leq 1)$ を求めよう.



ここまで来たよ



はじめに

- この授業どんなのり？



1 多次元確率分布

- 1 変数の連続型確率変数
- 2 次元の確率分布
- 周辺分布と独立性
- 離散-連続混合型

周辺分布

定義 (周辺分布)

$f(x, y)$: 連続型確率変数 X, Y の同時確率密度関数とする
 X, Y の周辺分布の確率密度関数は次で与えられる.

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

確率変数の独立性

Hoff ベイズ p.29

定義 (独立)

X, Y が独立とは, $f(x, y) = f_X(x) \times f_Y(y)$ と積で書けること.

命題 (独立なときの周辺分布)

X, Y が独立なとき,

- 1 変数の確率密度関数, 累積分布関数で考えられる.

$$\begin{aligned} P(a < X \leq b \text{ かつ } c < Y \leq d) &= \int_a^b dx \int_c^d dy f(x, y) \\ &= \int_a^b f_X(x) dx \times \int_c^d f_Y(y) dy \\ &= P(a < X \leq b) \times P(c < Y \leq d) \end{aligned}$$

L01-Q3

Quiz(2次元の確率密度関数)

連続型確率変数 X, Y の確率密度関数は次で与えられる.

$$f(x, y) = \begin{cases} 6e^{-2x-3y} & (x > 0, y > 0) \\ 0 & (\text{他}) \end{cases}$$

- ① 確率 $P(X \leq 1 \text{ かつ } Y \leq 5)$ を求めよう.
- ② 母期待値 $E[e^{-x+2y}]$ を求めよう.

L01-Q4

Quiz(2次元の確率密度関数)

連続型確率変数 X, Y の確率密度関数は次で与えられる.

$$f(x, y) = \begin{cases} 6e^{-2x-3y} & (x > 0, y > 0) \\ 0 & (\text{他}) \end{cases}$$

- ① 確率 $P(X \leq 1 \text{ または } Y \leq 5)$ を求めよう.
- ② 確率 $P(Y \leq 5X)$ を求めよう.
- ③ 母期待値 $E[3 + e^{x+y}]$ を求めよう.

ここまで来たよ



はじめに

- この授業どんなのり？



1 多次元確率分布

- 1 変数の連続型確率変数
- 2 次元の確率分布
- 周辺分布と独立性
- 離散-連続混合型

離散-連続混合型

Hoff ベイズ p.28

確率変数の組 (X, Y) を考える.

X : 連続型確率変数,

Y : 離散型確率変数, $Y = y_1, \dots, y_n$.

次を満たす確率密度関数 $f(x, y)$ で指定される.

定義 (離散-連続混合型分布)

$$P(a < X \leq b, Y = y_0) = \int_a^b f(x', y_0) dx'.$$

$$P(a < X \leq b, y_i \leq Y \leq y_j) = \int_a^b \sum_{y_i \leq y' \leq y_j} f(x', y') dx'$$

累積分布関数

$$F(x, y) := \int_{-\infty}^x \sum_{y'=-\infty}^y f(x', y') dx'.$$

気分を出すために, 表や場合分けで $f(x, y)$. ただし, $Y = 0, 1$. x についての確率密度関数 $h_0(x), h_1(x)$.

$$f(x, y) = \begin{cases} h_0(x) & (y = 0) \\ h_1(x) & (y = 1) \\ 0 & (y \neq 0, 1) \end{cases}$$

$y \backslash x$	$-\infty < x < +\infty$
0	$h_0(x)$
1	$h_1(x)$

$$\text{全} = \int_{-\infty}^{+\infty} h_0(x) \, dx + \int_{-\infty}^{+\infty} h_1(x) \, dx = 1.$$

混合ガウス分布 GMM=Gaussian mixture model

正規分布 normal distribution = ガウス分布 Gaussian distribution

岩薩林 確率・統計 §4.5

混合する = mix

混合ガウス分布を定義する準備として, $\pi_0 + \pi_1 = 1, \pi_y \geq 0, \pi_{\text{他}} = 0$ として, 次の同時分布 $f(x, y)$ を考える.

$$f(x, y) = \pi_y \frac{1}{(2\pi\sigma_y^2)^{1/2}} e^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} = \begin{cases} \pi_0 \cdot \frac{1}{(2\pi\sigma_0^2)^{1/2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} & (y = 0) \\ \pi_1 \cdot \frac{1}{(2\pi\sigma_1^2)^{1/2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} & (y = 1) \\ 0 & (y \neq 0, 1) \end{cases}$$

$$= \delta_{y0} \pi_0 \frac{1}{(2\pi\sigma_0^2)^{1/2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} + \delta_{y1} \pi_1 \frac{1}{(2\pi\sigma_1^2)^{1/2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}.$$

L01-Q5

Quiz(混合ガウス分布の確率密度関数)

X は混合ガウス分布 ($\pi_0 = \frac{3}{10}, \pi_1 = \frac{7}{10}, \mu_0 = 2, \mu_1 = 6, \sigma_0 = 2, \sigma_1 = \frac{1}{2}$) にしたがう. 確率密度関数 $f(x)$ のグラフの概形を描こう.
手で, または, `scipy.stats.norm` の浮動小数点計算で.

Y の周辺分布は離散型でベルヌイ分布 $B(1, \pi_1)$

$$f_Y(y) = \int f(x, y) dx = \int_{-\infty}^{+\infty} \pi_y \cdot \frac{1}{(2\pi\sigma_y^2)^{1/2}} e^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} dx = \pi_y.$$

X の周辺分布は連続型で GMM

定義 (GMM=混合ガウス分布)

連続型確率変数 X がパラメタ $(\pi_0, \pi_1, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2)$ の GMM に従うとは、次の確率密度関数にしたがうことをいう。

$$f_X(x) = \sum_y f(x, y) = \pi_0 \cdot \frac{1}{(2\pi\sigma_0^2)^{1/2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} + \pi_1 \cdot \frac{1}{(2\pi\sigma_1^2)^{1/2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}.$$

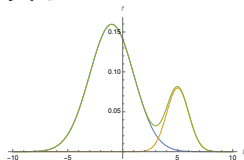
一般の混合ガウス分布: y の値が 2 個と限らず有限個

`scipy.stats` には GMM はないので, `scipy.stats.norm` を使って自分で計算する.

`sklearn.mixture.GaussianMixture`

一般の混合分布: ガウス分布とは限らない分布を重み π_y で「重ね合わせ」

注: 混合 $\pi_0 h_0(x) + \pi_1 h_1(x)$ と, 確率関数の和 $S = X_1 + X_2$ はもちろん異なる.



L01-Q6

Quiz(混合ガウス分布の確率密度関数)

X は混合ガウス分布 ($\pi_0 = \frac{3}{10}, \pi_1 = \frac{7}{10}, \mu_0 = 2, \mu_1 = 6, \sigma_0 = 2, \sigma_1 = \frac{1}{2}$) にしたがう. 確率密度関数 $f(x)$ のグラフの概形を描こう.
手で, または, `scipy.stats.norm` の浮動小数点計算で.

L01-Q7

Quiz(混合ガウス分布の確率)

X は混合ガウス分布 ($\pi_0 = \frac{3}{10}, \pi_1 = \frac{7}{10}, \mu_0 = 2, \mu_1 = 6, \sigma_0 = 2, \sigma_1 = \frac{1}{2}$) にしたがる.

- ① X の確率密度関数の $x = 4$ における値を求めよう.
- ② 確率 $P(X \leq 4)$ を, 標準正規分布の累積分布関数 $\Phi(x)$ で表そう.

さらに `scipy.stats.norm` で浮動小数点数で求めてみると安心できるかも.