クレジット:
UTokyo Online Education 統計データ解析 Ⅱ 2018 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析 (II) 第 6 回

小池祐太

2018 年 5 月 24 日

1. 連続分布
   - 正規分布
   - 一様分布
   - ガンマ分布
   - t 分布
   - F 分布

2. 多次元確率変数と多変量分布
   - 多項分布
   - 多変量正規分布

確率変数と確率分布

- **確率変数 (random variable)**
  - 値がランダムに決定される変数で, すべての実数 $a \leq b$ に対して, その値が区間 $[a, b]$ に含まれる確率があらかじめ定められているような変数 [1]
- X を確率変数とすると, 定義より X が区間 $[a, b]$ $(a \leq b)$ に含まれる確率が定まるから, その確率を

$$P(a \leq X \leq b)$$

  で表す
- 特に $a = b$ のとき, $P(a \leq X \leq b)$ は $X = a$ となる確率を表すから, それを $P(X = a)$ で表す

[1] この定義は数学的には厳密性を欠くが, 本講義ではこの定義を採用する.

確率変数と確率分布

- 一口に「値がランダムに決定される」といっても, 出現しやすい数値や, まったく出現しない数値があるかもしれない
- 確率統計学ではこのような値の出現頻度 (確率) を決定する法則が確率変数の背後に存在すると考えて, その法則を**確率分布 (probability distribution)** または単に**分布**と呼び, 確率分布の数学的モデリングを通じて現象の理解を試みる
- 確率分布の定義をもう少し正確に述べると, 確率変数 X に対して, 各区間 $[a, b]$ $(a \leq b)$ と, X が区間 $[a, b]$ に含まれる確率

$$P(a \leq X \leq b)$$

との対応を示したものを, X の確率分布または単に分布という.[2]

- また, このとき X はこの分布に**従う**という

[2] より現代的な定義を述べるためには測度論の知識が必要となるため, ここでは簡易的な定義を述べた.

連続分布

- 実際のデータでは, 取りうる値が任意の実数またはある範囲の実数である場合, もしくは取りうる値のパターンが数多いため近似的にすべての実数値またはある範囲の実数値をとりうると考えられる場合が頻繁にある
  - 具体例: 株価, 気温, 風速, 液体の体積など
- このようなデータのモデル化には, しばしば連続分布に従う確率変数が用いられる
- さらに, 以下で見るように, 離散分布に従うデータであっても, サンプル数が非常に大きい状況ではその分布はしばしば連続分布で近似できる
- このように, 離散的なデータの解析であったとしても, 連続分布を考えることは理論上重要となる

連続分布

- 一般に, 確率変数 X が**連続型 (continuous)** であるとは, 非負の値をとる実数直線上の関数 f があって, $a \leq b$ なるすべての実数 a, b に対して

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

  が成り立つことをいい, 対応する確率分布を**連続分布**と呼ぶ
- また, 関数 f をこの確率分布の**確率密度関数 (probability density function)**, あるいは単に**密度 (density)** と呼ぶ

連続分布

- 確率変数 X をシミュレーションした際のヒストグラムのビン $[a, b]$ における高さは

$$\frac{1}{b-a} P(a \leq X \leq b)$$

で与えられる (関数 `hist()` でオプション `freq` を `FALSE` に指定した場合)

- 従って, 確率密度関数 f は, ビン $[a, b]$ の幅を限りなく小さくした場合のヒストグラムの形状の極限として現れるグラフに対応する

正規分布

- μ を実数, σ を正の実数とする
- 確率密度関数が

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

で与えられる連続分布を平均 μ, 分散 σ^2 の**正規分布 (normal distribution)** または **Gauss 分布**と呼び, 記号 $N(\mu, \sigma^2)$ で表す

- 特に, 平均 0, 分散 1 の正規分布を**標準正規分布 (standard normal distribution)** と呼ぶ

正規分布

- ここで, 「平均」, 「分散」, 「標準偏差」という言葉は, データから計算される平均, 分散, 標準偏差とは意味合いが異なることに注意する必要がある
- 両者を区別するために, 後者の文頭に「標本」という言葉をつける場合がある
- 適当な仮定のもとで, データ数が大きくなるにつれて, 後者の意味での平均, 分散, 標準偏差はそれぞれ前者の意味での値に近づいていくことが知られている (大数の法則)

正規分布

- 正規分布に従う乱数の発生には関数 `rnorm( )` を用いる
- なお, 連続分布の場合, 分布の省略形の文頭に `d` をつけることで, 確率密度関数を計算するための関数が得られる
- 例えば, 正規分布の確率密度関数は関数 `dnorm( )` で計算できる
- 実行例 `rnorm2.r`

正規分布

- 正規分布は離散分布の極限としても現れる
- Y を試行回数 n, 成功確率 p の二項分布に従う確率変数とすると, n が十分大きいとき, $(Y - np)/\sqrt{np(1-p)}$ の分布は標準正規分布で近似できる
- これは **de Moivre-Laplace の定理**として知られている. **中心極限定理 (central limit theorem)** はその一般化 (「統計データ解析 I」の講義ノート 5 章参照)
- 実行例 `rbinom-normal2.r`

一様分布

- $a < b$ とする
- 確率密度関数が

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \text{ のとき}, \\ 0 & \text{上記以外のとき} \end{cases}$$

で与えられる連続分布を区間 (a, b) 上の**一様分布 (uniform distribution)** と呼び, 記号 $U(a, b)$ で表す

- 一様分布に従う乱数の発生には関数 `runif()` を用いる
- 実行例 `runif2.r`

ガンマ分布

- ν, α を正の実数とする
- 確率密度関数が

$$f(x) = \frac{1}{\Gamma(\nu)} \alpha^{\nu} x^{\nu-1} e^{-\alpha x} \quad (x > 0), \qquad f(x) = 0 \quad (x \leq 0)$$

で与えられる連続分布をパラメータ ν, α の**ガンマ分布 (gamma distribution)** と呼び，記号 $\Gamma(\nu, \alpha)$ や $G(\alpha, \nu)$ で表す

  - $\Gamma(\nu)$ は**ガンマ関数 (gamma function)**

$$\Gamma(\nu) = \int_0^{\infty} x^{\nu-1} e^{-x} dx$$

を表す

ガンマ分布

- ν, α はそれぞれ**形状パラメーター (shape)**, **レート (rate)** と呼ばれることがある
- ガンマ分布に従う乱数の発生には関数 `rgamma()` を用いる
- 実行例 `rgamma2.r`

指数分布

- ガンマ分布はいくつかの応用上重要な確率分布を特殊な場合として含む
- 正の実数 λ に対して, $\Gamma(1, \lambda)$ をパラメータ λ の**指数分布 (exponential distribution)** と呼び, 記号 $\mathrm{Exp}(\lambda)$ で表す
- λ は**レート**と呼ばれることがある
- 指数分布に従う乱数の発生には関数 `rexp()` を用いる
- 実行例 `rexp.r`

χ^2 分布

- 正の実数 k に対して, $\Gamma(k/2, 1/2)$ を自由度 k の χ^2 **分布**と呼び, 記号 $\chi^2(k)$ で表す [3]
- χ^2 分布に従う乱数の発生には関数 `rchisq()` を用いる
- 実行例 `rchisq.r`

[3] χ^2 は「カイ二乗」と読む

χ^2 分布

- 標準正規分布に従う k 個の独立な確率変数の二乗和は自由度 k の χ^2 分布に従うことが知られている
- この事実は推定や検定の理論において重要な役割を果たす (「統計データ解析 I」講義ノート 8–9 章参照)
- 実行例 `rgamma-chi2.r`

t 分布

- ν を正の実数とする
- 確率密度関数が

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

で与えられる連続分布を, 自由度 ν の (Student の) t **分布**と呼び, 記号 $t(\nu)$ で表す [4]

- t 分布に従う乱数の発生には関数 `rt( )` を用いる

[4] Student は t 分布を導入した統計学者 Gosset のペンネームである

t 分布

- Z を標準正規分布に従う確率変数, Y を自由度 k の χ^2 分布に従う確率変数とし, Z, Y は独立であるとする. このとき, 確率変数

$$\frac{Z}{\sqrt{Y/k}}$$

  は自由度 k の t 分布に従うことが知られている
- 実行例 `rt.r`

F 分布

- ν_1, ν_2 を正の実数とする
- 確率密度関数が

$$f(x) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2, \nu_2/2)} \frac{x^{\nu_1/2-1}}{(1+\nu_1 x/\nu_2)^{(\nu_1+\nu_2)/2}} \qquad (x > 0),$$
$$= 0 \qquad (x \leq 0)$$

で与えられる連続分布を, 自由度 ν_1, ν_2 の F **分布**と呼び, 記号 $F(\nu_1, \nu_2)$ で表す

- F 分布に従う乱数の発生には関数 `rf()` を用いる

F 分布

- Y_1 を自由度 k_1 の χ^2 分布に従う確率変数, Y_2 を自由度 k_2 の χ^2 分布に従う確率変数とし, Y_1, Y_2 は独立であるとする
- このとき, 確率変数

$$\frac{Y_1/k_1}{Y_2/k_2}$$

は自由度 k_1, k_2 の F 分布に従うことが知られている
- 実行例 `rf.r`

多次元確率変数と多変量分布

- 本講義では多変量データを扱うので, 確率変数の多次元版を考える必要がある
- 値がランダムに決定される d 次元ベクトルで, 各座標が確率変数であるようなものを d **次元確率変数 (**d**-dimensional random variable)** または d **次元確率ベクトル (**d**-dimensional random vector)** と呼ぶ.[5]

[5] 以下特に断らない限り, ベクトルは列ベクトルとみなす.

多次元確率変数と多変量分布

- 1 次元の場合の多次元化として, 多次元確率変数の分布を以下のようにして定義する
- d 次元確率変数 $X = (X_1, X_2, \ldots, X_d)^\top$ に対して, d 次元長方形 $\{(x_1, \ldots, x_d) : a_i \leq x_i \leq b_i \ (i = 1, \ldots, d)\}$ $(a_i \leq b_i,\ i = 1, \ldots, d)$ と, X がこの d 次元長方形に含まれる確率

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \ldots, a_d \leq X_d \leq b_d)$$

との対応を示したものを, X の **(d 変量) 確率分布**または単に **(d 変量) 分布**といい, X はこの分布に**従う**という

多項分布

- k を正整数とする
- 1 回の試行で起こり得る k 個の排反な事象 $E_1, \ldots, E_k$ があり，どれかは必ず起こるとする．E_i の起こる確率が p_i であるとする
- この試行を独立に n 回繰り返し，E_i が起こった回数を X_i とする
- このようにして定義される k 次元確率変数 $X = (X_1, \ldots, X_k)^\top$ の分布を試行回数 n, 確率 $p_1, \ldots, p_k$ の k **項分布**と呼ぶ
- 総称して**多項分布 (multinomial distribution)** と呼ぶ
- なお, 各 $i = 1, \ldots, k$ に対して, 確率変数 X_i の分布は試行回数 n, 成功確率 p_i の二項分布であることが確認できる
  - この意味で, 多項分布は二項分布の多変量版であると考えられる

多項分布

- いまの場合, k 次元確率変数 X は有限個の値しかとりえないため, 1 次元の場合と同様その分布は X のとりうる値 x のそれぞれに対して X が値 x をとる確率を対応させる関数 $f(x)$ を与えることで完全に決定される
- この関数 f を X(の分布) の**確率 (質量) 関数**と呼ぶ
- 試行回数 n, 確率 $p_1, \ldots, p_k$ の k 項分布の確率関数は, $0 \leq x_i \leq n$ $(i = 1, \ldots, k)$, $\sum_{i=1}^{k} x_i = n$ なる整数の組 $(x_1, \ldots, x_k)$ に対して定義され,

$$f(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

で与えられる

多変量の乱数

- 1 次元の場合と同様にして, d 次元確率変数の列の独立性が定義される (区間を d 次元長方形に置き換えればよい)
- 1 次元の場合と同様に, 独立な多次元確率変数列を (多変量の) 乱数と呼ぶことにする
- 1 変量分布の場合と異なり, 多変量分布に従う乱数の生成は容易でないことが多く, R にデフォルトで実装されている関数もほとんどない
- 多項分布は数少ない例外であり, 関数 `rmultinom()` によって多項分布に従う乱数を生成できる
- 実行例 `rmultinom.r`

多変量正規分布

- 1 変量の場合と同様に, 多変量の場合も連続分布が定義される
- d 次元確率変数 X(の分布) が**連続型**であるとは, ある d 個の変数をもつ非負値関数 $f(x_1, \ldots, x_d)$ が存在して, $a_i \leq b_i$ $(i = 1, 2, \ldots, d)$ なる任意の実数 $a_1, b_1, a_2, b_2, \ldots, a_d, b_d$ に対して,

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \ldots, a_d \leq X_d \leq b_d) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_d}^{b_d} f(x_1, x_2, \ldots, x_d) dx_1 dx_2 \cdots dx_d$$

  が成り立つことをいう
- 1 変量の場合と同様 X の分布は関数 f によって完全に決定されることが知られており, この関数 f を X(の分布) の **(確率) 密度 (関数)** と呼ぶ

多変量正規分布

- 多変量連続分布の最も重要な例は多変量正規分布である
- $\boldsymbol{\mu}$ を d 次元ベクトル, Σ を d 次正定値対称行列[6] とするとき, 確率密度関数が

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

で与えられる連続型 d 変量分布を, 平均ベクトル $\boldsymbol{\mu}$, 共分散行列 Σ の d **変量正規分布 (d-dimensional normal distribution)** と呼ぶ

  - Σ の行列式は Σ の固有値の積で与えられるから, $\det \Sigma > 0$ であり, 特に Σ は正則である
  - d 次元ベクトル $\boldsymbol{x}$ は列ベクトルとみなしている

[6] d 次対称行列 Σ が**正定値 (positive definite)** であるとは, Σ の固有値がすべて正であることをいう.

多変量正規分布

- 平均ベクトルが零ベクトルで共分散行列が単位行列の d 変量正規分布を d **変量標準正規分布 (**d**-dimensional standard normal distribution)** と呼ぶ
- R にはデフォルトでは多変量正規分布に従う乱数を生成するための関数は用意されていないため, 自作する必要がある
  - パッケージをインストールする方法もある (後述)
- 多変量正規分布のシミュレーションを行うためには, 次の命題が有用である:

多変量正規分布

命題 1

(a) $X_1, \ldots, X_d$ を標準正規分布に従う独立な d 個の確率変数とする．このとき，d 次元確率変数 $X = (X_1, \ldots, X_d)^\top$ は d 変量標準正規分布に従う．
(b) X を平均ベクトル $\boldsymbol{\mu}$, 共分散行列 Σ の d 変量正規分布に従う d 変量確率変数とする．このとき，A が d 次正則行列，$\boldsymbol{b}$ が d 次元列ベクトルならば，d 次元確率変数 $AX + \boldsymbol{b}$ は平均ベクトル $A\boldsymbol{\mu} + \boldsymbol{b}$, 共分散行列 $A\Sigma A^\top$ の d 変量正規分布に従う．

多変量正規分布

- 命題 1 より, d 次元ベクトル $\boldsymbol{\mu}$ および d 次元正定値対称行列 Σ が与えられたとき, 以下の手順によって平均ベクトル $\boldsymbol{\mu}$, 共分散行列 Σ の d 変量正規分布に従う d 次元確率変数 X を生成できる:
  (1) d 個の標準正規乱数 $Z_1, \ldots, Z_d$ を生成し, $Z = (Z_1, \ldots, Z_d)^\top$ とおく. 命題 1(a) より Z は d 変量標準正規分布に従う.
  (2) d 次正方行列 A で $\Sigma = AA^\top$ を満たすものを計算し, $X = \boldsymbol{\mu} + AZ$ とおく. 命題 1(b) より X は平均ベクトル $\boldsymbol{\mu}$, 共分散行列 Σ の d 変量正規分布に従う.
- 上の手順のうち, (1) における標準正規乱数の生成は関数 `rnorm( )` によって実行できる
- 手順 (2) における行列 A の計算にはいくつか方法があるが, ここでは固有値分解を用いる方法を説明する.[7]

[7] 別の方法としては, 例えば配布資料 2.7.3 節で説明したコレスキー分解を使う方法があり, より直接的である.

多変量正規分布

- Σ は対称行列だから, ある d 次正則行列 V によって対角化できる:

$$V^{-1}\Sigma V = \Lambda$$

  - Λ は Σ の固有値を対角成分とする対角行列
- さらに, V を直交行列, すなわち $V^{-1} = V^{\top}$ となるようにとることができることが知られている
- いま, Σ は正定値であったから, Λ の対角成分 $\lambda_1, \ldots, \lambda_d$ はすべて正
- 従って, $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}$ を対角成分とする対角行列 D を考えることができる

多変量正規分布

- このとき $A := VDV^\top$ とおくと,

$$AA^\top = VDV^\top (VDV^\top)^\top = VDV^\top VDV^\top = VD^2V^\top = V\Lambda V^\top = \Sigma$$

  となるので, この行列 A が求めるべきものである
- 多変量正規分布に従う乱数を発生させるための関数を実装しているパッケージはいくつか存在する
  - パッケージ `MASS` には関数 `mvrnorm()` が, パッケージ `mvtnorm` には関数 `rmvnorm()` がそれぞれ多変量正規分布に従う乱数を発生させるための関数として実装されている
- 実行例 `rmvnorm2.r`