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 5 回

小池祐太

2018 年 5 月 17 日

1. データの可視化
   - 基本的な描画
   - ヒストグラム
   - 箱ひげ図
   - 相関行列の可視化
   - その他の描画関数
2. 乱数
3. 確率変数と確率分布
4. 離散分布
   - 二項分布
   - Poisson 分布
5. 連続分布
   - 正規分布
   - 一様分布
   - ガンマ分布
   - t 分布
   - F 分布

データのプロット

- データ全体の特徴や傾向を把握するために効果的な方法は, データの可視化である
- R にはきわめて多彩な作図機能が用意されており, ここではいくつかの代表的な描画関数を取り上げて解説する
- 描画関連の関数は色, 線種や線の太さ, あるいは図中の文字の大きさなどを指定するために, 多彩なオプションを用意しており, ここでは説明しきれないため, 必要に応じて関数 `help()`(ヘルプの表示) と `example()`(例題の表示) を参照

基本的な描画 (日本語を含む図の描画)

- OS によっては日本語を含む図を描画すると文字化けする場合がある
- その場合, 関数 `par()` のオプション `family` に適当なフォントファミリーを指定することで文字化けを回避できる場合がある
- 例えば, Mac OS のデフォルトの設定では日本語を含む図は文字化けしてしまうが, 以下のコマンドをコンソール上で実行することで文字化けを回避できる

```
par(family = "HiraginoSans-W4")
```

  - フォントファミリーとしてヒラギノ角ゴシック W4 を指定している (数字を変えると太さが変わる)

- 実行例 `plot-kion.r`

ヒストグラム

- **ヒストグラム (histogram)**
  - データの値の範囲をいくつかの区間に分割し, 各区間に含まれるデータの個数を棒グラフにしたもの
  - 棒グラフの横幅が区間に対応し, 面積が区間に含まれるデータの個数に比例するようにグラフを作成する
  - データの分布の仕方 (どのあたりに値が集中しているか, どの程度値にばらつきがあるかなど) を可視化するのに有効
- ヒストグラムは関数 `hist()` で作成できる

ヒストグラム

- 基本書式

`hist(x, breaks, freq)`

  - `x`: ヒストグラムを描画するベクトル
  - `breaks`: 区間の分割の仕方を指定. 数字を指定するとデータ範囲をその数字に近い個数に等分割する. デフォルトの個数は Sturges の公式によって決定される. すなわち, データ数を n とすると, $\lceil \log_2 n + 1 \rceil$ である.[1] その他の指定方法もある (ヘルプ参照)
  - `freq`: `TRUE` 指定すると縦軸をデータ数にし, `FALSE` 指定すると縦軸をデータ数/全データ数とする. デフォルトは `TRUE` (`breaks` の指定によって変わる場合あり)
  - 他にも `plot` で指定できるオプションが利用可能

- 実行例 `hist3.r`

[1] $\lceil x \rceil$ は x 以下の最大の整数を表す.

箱ひげ図

- **箱ひげ図 (boxplot)**
  - データの中心, 散らばり具合および外れ値を考察するための図 (ヒストグラムの簡易版)
  - 複数のデータの分布の比較の際に有効
  - データの第 1 四分位点を下端, 第 3 四分位点を上端とする長方形 (箱) と, 第 1 四分位点, 第 3 四分位点からそれぞれ箱の長さの 1.5 倍以内にあるデータのうちの最小の値, 最大の値を下端, 上端とする直線 (ひげ) からなる
  - ひげの外側のデータは点で表示される
  - 中央値は太線で表示される
- 箱ひげ図は関数 `boxplot()` で描画できる

箱ひげ図

- ベクトル x に対する箱ひげ図は boxplot(x,...) で描画できる (... に関数 plot( ) と同様のオプションを指定可能)
- データフレーム x に対して, boxplot(x,...) は列ごとの箱ひげ図を描画
- データフレーム x において, 変数 A が「分類」を表す変数 (性別, 植物の種類など)[2] の場合, 別の変数 B に対して,

```
boxplot(B ~ A, data = x, ...)
```

は変数 B を変数 A で分類した場合の, 分類ごとの箱ひげ図を描画する

- 実行例 boxplot2.r

[2] **質的変数**と呼ばれる

相関行列の可視化

- 列数が非常に多い大規模データフレームの変数間の相関の様子を見る場合, 相関行列の可視化が便利である
- パッケージ `corrplot` には相関行列を可視化するための関数 `corrplot()` および関数 `corrplot.mixed()` が用意されている
- 実行例 `corrplot.r`

その他の描画関数

- データフレーム x に対して plot(x,...) もしくは pairs(x, ...) を実行すると, すべての列のペアに対する散布図を行列状に並べた図を作成する
  - 変数 A1, ..., Ak のみ考えたい場合, plot(~ A1 + ··· + Ak, data = x, ...) もしくは pairs(~ A1 + ··· + Ak, data = x, ...) を利用
- ベクトル x の各成分の値に基づく円グラフの作成は, pie(x, ...) で実行できる
- 実行例 graphic-misc2.r

乱数

- **乱数 (random numbers)**
  - ランダムに生成された数列
  - もちろん, コンピューターでは完全にランダムに数字を発生されることは不可能なため, コンピューター上で発生された乱数はすべて**擬似乱数 (pseudo random numbers)** である
    - ★ R では擬似乱数を発生させるための方法として Mersenne ツイスターがデフォルトでは用いられている (`help(Random)` 参照)
  - 特に, 数値シミュレーションを行う上では, それが再現可能であることが要請されるため, 発生される乱数も再現可能である必要がある
  - R ではこれを実行するために, 乱数の初期値を指定するための関数 `set.seed()` が用意されている (同一の初期値から生成される乱数は同一のものとなる)
- 実行例 `sample.r`

確率変数と確率分布

- 数学的には, 乱数は**確率変数 (random variable)** という概念でモデル化される
  - 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 [3]
- 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

[3] 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

- 乱数はランダムに生成された数列であったが, この場合「ランダム」という言葉は以下の 2 種類の意味に使われている:
  - (i) 数列の個々の数字がランダムに決定されている.
  - (ii) 数列の値の並び方に規則性がない (個々の数字がとる値が他の数字がとる値に影響しない).
- (i) のランダム性は確率変数によってモデル化できる
- (ii) のランダム性は, 数学的には**独立性 (independence)** という概念でモデル化される

確率変数と確率分布

- 確率変数の列 $X_1, X_2, \ldots, X_n$ が**独立 (independent)** であるとは, $a_i \leq b_i$ $(i = 1, \ldots, n)$ なる任意の実数 $a_1, b_1, \ldots, a_n, b_n$ に対して,

$$\begin{aligned} &P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \ldots, a_n \leq X_n \leq b_n) \\ &\quad = P(a_1 \leq X_1 \leq b_1)P(a_2 \leq X_2 \leq b_2) \cdots P(a_n \leq X_n \leq b_n) \quad (1) \end{aligned}$$

  が成り立つことをいう
  - (1) の左辺は「X_1 が区間 $[a_1, b_1]$ に値をとり, X_2 が区間 $[a_2, b_2]$ に値をとり, $\ldots$, X_n が区間 $[a_n, b_n]$ に値をとる」という事象が起きる確率を表す
- 従って, 以下で乱数というときは, 数学的には独立な確率変数列を指すものとする

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.[4]

- また, このとき X はこの分布に**従う**という

[4] 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.

離散分布

- 取りうる値が有限個, もしくは可算無限個 (例えば整数値のみとる場合) であるような確率変数は**離散型 (discrete)** であるといい, 対応する確率分布を**離散分布**と呼ぶ
- 離散分布は, その分布に従う確率変数 X が取りうる値 x のそれぞれに対して, $X = x$ となる確率 $P(X = x)$ を対応させる関数 $f(x) = P(X = x)$ を考えることで完全に決定される
- この関数 f を**確率質量関数 (probability mass function)**, あるいは単に**確率関数 (probability function)** と呼ぶ

二項分布

- n を正の整数, p を 0 以上 1 以下の実数とする
- 取りうる値が $0, 1, \ldots, n$ であり, 確率関数が

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n$$

  で与えられる離散分布を, 試行回数 n, 成功確率 p の**二項分布 (binomial distribution)** と呼ぶ
- 特に, 試行回数 1 の二項分布を **Bernoulli 分布 (Bernoulli distribution)** と呼ぶ
- 例えば, 表が出る確率が p のコインを n 回投げたときに表が出る回数は試行回数 n, 成功確率 p の二項分布に従う

二項分布

- 二項分布に従う乱数の発生には関数 `rbinom( )` を用いる
- なお, 原則として, ある確率分布に従う乱数を生成するための R の関数の命名規則は,「`r` + その乱数が従う分布の名前の省略形」となっている (一部例外がある)
- また, 離散分布の場合, その確率関数を計算するための関数が, 同じ省略形の文頭に `d` をつけることで得られる
- 例えば, 二項分布の確率関数は関数 `dbinom( )` で計算できる
- 実行例 `rbinom2.r`

Poisson 分布

- λ を正の実数とする
- 取りうる値が 0 以上の整数であり, 確率関数が

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, \ldots$$

で与えられる離散分布をパラメーター λ の **Poisson 分布 (Poisson distribution)** と呼び, 記号 $P_o(\lambda)$ で表す

  - λ は**強度 (intensity)** と呼ばれることがある
  - 放射性物質から一定時間に放射される粒子の数や, 一定期間に起こる交通事故の数などは Poisson 分布に従うことが知られている
  - 発生確率が低い事象が十分長い期間のあいだに起こる回数の分布は Poisson 分布で近似できる (少数の法則)

- Poisson 分布に従う乱数の発生には関数 `rpois()` を用いる
- 実行例 `rpois2.r`

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 [5]
- 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`

[5] 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 [6]

- t 分布に従う乱数の発生には関数 `rt( )` を用いる
- 実行例 `rt2.r`

[6] 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 `normal-t.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
- 実行例 `rf2.r`

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 `normal-f.r`