クレジット:
UTokyo Online Education 統計データ解析 I 2017 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析 (I) 第 6 回

小池祐太

2017 年 11 月 1 日

3 次元のグラフ

- 様々な種類の 3 次元のグラフを描画するために多くのパッケージが開発されている
- そのうちの 1 つパッケージ `scatterplot3d` には, 3 次元の散布図を書くための関数 `scatterplot3d()` が用意されている
- 基本書式

```
scatterplot3d(x, color, angle=40)
```

  - `x`: x, y, z 座標を指定するデータフレーム (関数 `persp()` のように直接指定することも可能)
  - `color`: 色を指定 (`col` ではない). デフォルトは黒
  - `angle`: x 軸と y 軸の間の角度
  - 他にも `plot` で指定できるオプションが利用可能
- 実行例 `plot3d.r`

プロット環境の設定

- プロットの際の線の種類や色, 点の形等のデフォルト値は関数 `par( )` で設定できる
- 設定可能なグラフィックスパラメーターは `help(par)` で確認できる
- 特に, 以下の例のように, 関数 `par( )` によってプロット環境の設定 (複数図の配置, 余白の設定) ができる
- 実行例 `par.r`

プロット環境の設定

- プロット環境は非常に細かく設定でき, またそれぞれの描画関数独自のパラメーターも存在するため, ここでは紹介しきれない. 必要に応じてヘルプファイル, またはインターネット上の情報を参照すること

シミュレーションと極限定理

- データ解析の際, 分析対象としたい集団全体のデータが入手できることは多くの場合稀
- そもそも集団全体のデータの入手は不可能な場合もある
  - データ解析の目的が「将来の予測」や「集団の背後にある共通の法則の発見」であった場合, 分析対象としたい集団に「将来のデータ」や「(必ずしも現時点の集団に含まれているとは限らない) 実現する可能性のあるデータ」が含まれてしまうため
- このような理由から, 多くの場合, 分析対象の集団の一部のデータのみを対象としてデータ解析を実行し, その結果から集団全体の性質についての知見を得る必要が生じる
- そのための方法論を体系的に研究する分野を**推測統計**と呼ぶ

シミュレーションと極限定理

- 分析対象の集団の一部から解析のためのデータ収集を行う際, データの集め方に偏りがあると, データ解析の結果から集団全体の性質を推測することは難しくなることは直感的には明らか
  - 例えば, 日本全体の平均気温を計測したい場合に, 沖縄県の各地点の気温のみ計測してデータとしてしまうと, 得られたデータから計算された平均気温は明らかに真に知りたい値より高くなってしまう
- このような問題を回避するためには, データを「ランダムに」収集すれば良いことは, 直感的・経験的にはよく知られている
- しかし, 「ランダムに」データを収集することでなぜ問題が解決できるのかということの数学的・論理的な根拠は, 実際にはそれほど自明でない

シミュレーションと極限定理

- この問いに厳密な意味での解答を与えるためには, 数学の一分野である「(測度論的) 確率論」を学習する必要があるが, その理解のためには他の数学の分野にも習熟する必要がある
- 本講義では, そのような問題を避けて, ランダムネスによって上述のサンプリングの問題などのデータ解析上の困難が解決できることを直感的に理解するために, 乱数を使ったシミュレーションによって, ランダムネスから結論される種々の数学的結果を観察する

乱数

- **乱数**とはランダムに生成された数列のことである
- 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, それらの乱数は厳密には**擬似乱数**である.[1]
- 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)

[1] 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.

乱数

- ここでは基本的な乱数として, ランダムサンプリング, 二項乱数および一様乱数を考える
- **ランダムサンプリング**
  - 与えられた集合の要素をランダムに抽出することで発生する乱数
- **二項乱数**
  - 「確率 p で表がでるコインを n 回投げた際の表が出る回数」に対応する乱数
  - 従って p と n によって乱数の発生の仕方が変わるため, それを明示するために「確率 p に対する次数 n の二項乱数」とも呼ぶ

乱数

- **一様乱数**
  - ある決まった区間 (a, b) $(a < b)$ に含まれる数字からランダムに発生する乱数 [2]
  - 従って区間 (a, b) によって乱数の発生の仕方が変わるため, それを明示するために「区間 (a, b) 上の一様乱数」とも呼ぶ
- ランダムサンプリングは関数 `sample()` で実行できる
- 二項乱数および一様乱数はそれぞれ関数 `rbinom()` および `runif()` で発生させられる
- 実行例 `sample.r`
- R には他にも様々な種類の確率分布に従う乱数が実装されているが, それらについては次回以降で詳しく説明する

[2] (a, b) は a より大きく b より小さい実数全体からなる集合を表す.

確率変数

- 数学的には, 乱数は**確率変数**という概念でモデル化される
- 確率変数とは, 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のことをいう [3]
- X が確率変数ならば, 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 $P(a \leq X \leq b)$ で表すことにする
- 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 $P(X = a)$ で表すことにする
- 以下本章では, 記述の簡単のために主として有限個の値のみをとる確率変数のみ考える
  - 無限個の値, 特に連続的な値をとる確率変数については次回以降の講義で説明する

[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.

平均と分散

- X を (とりうる値が有限個である) 確率変数として, X の取りうる値を $x_1, x_2, \ldots, x_N$ とする
- このとき, 以下で定義される量

$$E[X] := \sum_{i=1}^{N} x_i P(X = x_i)$$

を X の**平均**もしくは**期待値**と呼ぶ

  - $E[X]$ は X の「理論上の平均値」に対応する量とみなせる

- **例** 「偶数が出る確率が奇数の出る確率の 2 倍あるようなサイコロを振ったときに出る目」として定義される確率変数を X とする
- $p = P(X = 1)$ とおくと,

$$p = P(X = 3) = P(X = 5),$$
$$2p = P(X = 2) = P(X = 4) = P(X = 6)$$

であり, また $\sum_{x=1}^{6} P(X = x) = 1$ であるから, $9p = 1$, 従って $p = 1/9$

- 故に, X の平均は

$$\begin{aligned} E[X] &= \sum_{x=1}^{6} xP(X = x) \\ &= 1 \cdot p + 2 \cdot 2p + 3 \cdot p + 4 \cdot 2p + 5 \cdot p + 6 \cdot 2p \\ &= 33p = \frac{11}{3} = 3.6666\ldots \end{aligned}$$

平均と分散

- より一般に, X の関数 $\varphi(X)$ に対して, $\varphi(X)$ の期待値を

$$E[\varphi(X)] := \sum_{i=1}^{N} \varphi(x_i) P(X = x_i)$$

  で定義する
- 例えば, 正の整数 p に対して

$$E[X^p] = \sum_{i=1}^{N} x_i^p P(X = x_i)$$

  である
  - この量を X の p 次の**モーメント**あるいは**積率**と呼ぶ

平均と分散

- 測定誤差の大きさによって統計的推定の精度は左右されるため, 確率変数の値のばらつき具合を定量化する指標が必要である
- そのような指標の 1 つとして, 平均からのばらつき具合を定量化した指標

$$\mathrm{Var}[X] := E[(X - E[X])^2] = \sum_{i=1}^{N} (x_i - E[X])^2 P(X = x_i)$$

があり, これを X の**分散**と呼ぶ

平均と分散

- 分散はもとの確率変数を二乗したスケールをもつため, 単位をあわせるために, 分散の平方根

$$\sqrt{\mathrm{Var}[X]}$$

を考えることもよくある. この量を X の**標準偏差**と呼ぶ

- 分散の計算には次の恒等式が便利である:

$$\mathrm{Var}[X] = E[X^2] - (E[X])^2 \tag{1}$$

- **例** 上で考えた「偶数が出る確率が奇数の出る確率の 2 倍あるようなサイコロを振ったときに出る目」として定義される確率変数 X の分散を計算してみる

$$\begin{aligned} E[X^2] &= \sum_{x=1}^{6} x^2 P(X = x) \\ &= 1^2 \cdot p + 2^2 \cdot 2p + 3^2 \cdot p + 4^2 \cdot 2p + 5^2 \cdot p + 6^2 \cdot 2p \\ &= 147p = \frac{29}{3} \end{aligned}$$

であるから, 公式 (1) より

$$\mathrm{Var}[X] = \frac{29}{3} - \frac{121}{9} = \frac{26}{9} = 2.8888\ldots$$

- 標準偏差は $\sqrt{26}/3 = 1.69967\cdots$

独立性と同分布性

- 統計の文脈では, 確率変数は観測データの一つ一つに対応する
- 統計学では多数の観測データを扱うため, 確率変数の列 $X_1, X_2, \ldots, X_n$ に対する考察が重要となる
- 以下, 「X_1 が x_1 という値をとり, X_2 が x_2 という値をとり, $\ldots$, X_n が x_n という値をとる」という事象が起きる確率を, 記号

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

で表す

独立性と同分布性

観測データが「ランダムに」サンプリングされた状況を表現するために, 次の概念を導入する

定義 1 (確率変数列の独立性)

確率変数列 $X_1, X_2, \ldots, X_n$ が**独立**であるとは, 任意の n 個の実数 $x_1, x_2, \ldots, x_n$ に対して

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n)$$

が成り立つことをいう.

- 直感的には, $X_1, X_2, \ldots, X_n$ が独立であるというのは, すべての $i = 1, \ldots, n$ について, X_i がとる値は $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ がとる値と無関係に定まるということである

独立性と同分布性

- 独立性と並んで重要な概念に, 確率変数列の同分布性がある
- これは, 観測データが同一の法則に従って生成された集団からサンプルされたということを数学的に表現した概念である

定義 2 (確率変数列の同分布性)

確率変数列 $X_1, X_2, \ldots, X_n$ が**同分布**であるとは, 任意の実数 x に対して

$$P(X_1 = x) = P(X_2 = x) = \cdots = P(X_n = x)$$

が成り立つことをいう.

- 独立かつ同分布な確率変数列を**独立同分布**もしくは **i.i.d.** であるという (i.i.d. は independent and identically distributed の略)

大数の法則

- 分析対象の集団の平均値を求めたい場合, 分析対象の一部を「ランダムに」収集したデータのみを用いて (標本) 平均を計算しても, データのサンプル数が十分大きければ, その平均値は集団全体の平均値に近い値であると考えられることは, 経験則としてよく知られている (例えば視聴率の調査などがそうである)
- また, 表裏の出方に偏りがないコインを繰り返し投げ続けると, 投げた回数に対する表が出た回数の割合は理論上の平均値 $\frac{1}{2}$ に近づいて行くことも, 経験則として知られている
- このように, 同一の法則に従って生成された (と仮定された) 集団に対して「ランダムな」観測を多数繰り返すと, 観測値の平均は「真の平均値」(集団全体の平均値や理論上の平均値のこと) に近づくことは経験上よく知られている

大数の法則

- この法則を数学的に定式化した定理は**大数の法則**として知られている
- 数学的には，「サンプル数が十分大きい」という状況を「サンプル数を無限大にしたときの極限」として定式化する
- 従って確率変数の無限列 $X_1, X_2, \ldots$ を考察する必要が生じる
- この場合の独立性および同分布性を次頁で再定義しておく

大数の法則

定義 3 (無限列の場合の独立性と同分布性)

(a) $X_1, X_2, \ldots$ が**独立**であるとは, 任意の正整数 n に対して $X_1, X_2, \ldots, X_n$ が独立であることをいう.
(b) $X_1, X_2, \ldots$ が**同分布**であるとは, 任意の正整数 n に対して $X_1, X_2, \ldots, X_n$ が同分布であることをいう.
(c) $X_1, X_2, \ldots$ が**独立同分布**もしくは **i.i.d.** であるとは, $X_1, X_2, \ldots$ が独立かつ同分布であることをいう.

大数の法則

以上の定義のもと, 大数の法則は以下のように述べられる.

定理 1 (大数の強法則)

$X_1, X_2, \ldots$ を独立同分布な確率変数列とし, その平均を μ とする. このとき, $X_1, \ldots, X_n$ の標本平均

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

が $n \to \infty$ のとき μ に収束する確率は 1 である (このことを, 「$\frac{1}{n} \sum_{i=1}^{n} X_i$ は $n \to \infty$ のとき μ に**概収束する**」という).

- 実行例 `LLN.r`

大数の法則

演習 1

大数の法則の実行例 `LLN.r` において, 乱数の初期値・出現確率の変更・乱数の変更などを行っても大数の法則が成立することを確認せよ. また, サンプル数の増減によって精度がどの程度変わるか観察せよ.

中心極限定理

- 前節で説明した大数の法則は, サンプル数 n を大きくするに従い標本平均 $\bar{X}_n$ が真の平均 μ に限りなく近づくことを保証している
- 言い換えると,「推定誤差」$\bar{X}_n - \mu$ は n を大きくすると限りなく 0 に近づく
- しかし, 実際に推定誤差 $\bar{X}_n - \mu$ がどの程度の大きさになるのか定量的に評価する手段は与えていない

中心極限定理

- 統計学では, 推定誤差がある区間 $[\alpha, \beta]$ に入る確率

$$P(\alpha \leq \bar{X}_n - \mu \leq \beta) \tag{2}$$

  を計算することによって推定誤差を定量的に評価する (詳しい手順については「推定」の講義で説明する)
- 確率 (2) の正確な計算は一般には困難であるが, サンプル数が十分大きい場合ある関数の定積分で近似できることが知られている
- このことを具体的に述べたのが次の**中心極限定理**である

中心極限定理

定理 2 (中心極限定理)

$X_1, X_2, \ldots$ を独立同分布な確率変数列とし, その平均を μ, 標準偏差を σ とする. このとき, すべての実数 $a < b$ に対して

$$P\left(a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx \qquad (n \to \infty)$$

が成り立つ.

中心極限定理

- 中心極限定理より, サンプル数 n が十分大きければ, 確率

$$P\left(a\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - \mu \leq b\frac{\sigma}{\sqrt{n}}\right)$$

は定積分

$$\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}dx \tag{3}$$

によって近似できる

- 積分 (3) の被積分関数 $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ は**標準正規密度 (関数)** と呼ばれており, R では関数 `dnorm()` で計算できる

中心極限定理

- 定積分 (3) 自体は関数 `pnorm( )` を用いてコマンド

  `pnorm(b)-pnorm(a)`

  で計算できる
- 補足
  - 詳細は次回以降の講義で説明するが, 各実数 $a \leq b$ について区間 $[a, b]$ に値が入る確率が定積分 (3) で与えられるような確率変数を**標準正規確率変数**と呼び, 標準正規確率変数の値の分布の仕方を**標準正規分布**と呼ぶ
  - 中心極限定理の意味するところは, X_i たちの分布が何であっても, サンプル数 n が十分大きければ, 標本平均を正規化した量の $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ の分布の仕方は標準正規確率分布で近似できるということである

中心極限定理

- 中心極限定理のシミュレーションによる確認は, ヒストグラムによる可視化を用いる方法がよく利用される
- これは, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ をシミュレーションした際のヒストグラムのビン $[a, b]$ における高さが

$$\frac{1}{b-a} P\left(a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq b \right)$$

で与えられることを利用する (関数 `hist()` でオプション `freq` を `FALSE` に指定した場合)

中心極限定理

- 従ってビンの幅 $b - a$ が十分小さければ, 中心極限定理が正しい限りビン $[a, b]$ におけるヒストグラムの高さは $\phi(a)$ で近似できるはずである
- 従って, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ のヒストグラムに標準正規密度 $\phi(x)$ を重ね書きすることで, 近似の度合いを評価できる
- 実行例 `CLT.r`

中心極限定理

演習 2

中心極限定理の実行例 `CLT.r` において, 乱数の初期値・出現確率の変更・乱数の変更などを行っても中心極限定理が成立することを確認せよ. また, サンプル数の増減によって精度がどの程度変わるか観察せよ.

重複対数の法則

定理 3 (重複対数の法則)

$X_1, X_2, \ldots$ を独立同分布な確率変数列とし, その平均を μ, 標準偏差を σ とする. このとき,

$$\limsup_{n\to\infty} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log\log n}} = 1 \quad \text{a.s.}, \tag{4}$$

$$\liminf_{n\to\infty} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log\log n}} = -1 \quad \text{a.s.} \tag{5}$$

が成り立つ. より一般に, 列

$$\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log\log n}} \right)_{n=3}^{\infty}$$

のある部分列の収束先となるような実数全体の集合を C とすると, C が閉区間 $[-1, 1]$ に一致する確率は 1 である.

重複対数の法則

- 定理 3 の後半の主張は **Hartman-Wintner の定理**として知られている
- 実行例 `LIL.r`

重複対数の法則

演習 3

重複対数の法則の実行例 `LIL.r` において, 乱数の初期値・出現確率の変更・乱数の変更などを行っても重複対数の法則が成立することを確認せよ. また, サンプル数の増減によって精度がどの程度変わるか観察せよ.

少数の法則

- 少数の法則とは, 滅多に起こらない事象が起こる回数の分布に関する法則である
- 例えば, ある製品の不良品率 p はとても小さいとする. 一日に n 個 (非常に多数とする) 生産するとき, 不良品は平均的には $\lambda = np$ 個発生するが, 日によって不良品の個数 S_n には多少のばらつきが生じる
- 従って S_n は確率変数であるが, S_n がとる値の確率法則は, 強度 λ の **Poisson 分布**で近似できることが知られている
- これを正確に述べたのが次の**少数の法則**である

少数の法則

定理 4 (少数の法則)

$X_1, X_2, \ldots, X_n$ を独立な確率変数列とし, 各 $i = 1, 2, \ldots, n$ について X_i は確率 $p_{n,i}$ で 1 を, 確率 $1 - p_{n,i}$ で 0 をとるとする:

$$P(X_i = 1) = p_{n,i}, \qquad P(X_i = 0) = 1 - p_{n,i} \qquad (i = 1, 2, \ldots, n).$$

このとき, ある正の実数 λ が存在して, $n \to \infty$ のとき

$$\max_{i=1,2,\ldots,n} p_{n,i} \to 0, \qquad \sum_{i=1}^{n} p_{n,i} \to \lambda$$

が成り立つならば, 任意の 0 以上の整数 k に対して

$$P\left(\sum_{i=1}^{n} X_i = k\right) \to e^{-\lambda} \frac{\lambda^k}{k!} \qquad (n \to \infty)$$

が成り立つ.

少数の法則

- 上の定理において, $\sum_{i=1}^{n} X_i$ が本小節冒頭の例の S_n に対応する
- 補足
  - 詳細は次章で説明するが, 取りうる値が 0 以上の整数全体で, 値が整数 $k \geq 0$ となる確率が

$$e^{-\lambda} \frac{\lambda^k}{k!} \tag{6}$$

    で与えられる確率変数を強度 λ の **Poisson 型確率変数**と呼び, その値の分布の仕方を強度 λ の Poisson 分布と呼ぶ
  - (6) は関数 `dpois()` で計算できる
- 実行例 `LRE.r`

少数の法則

演習 4

少数の法則の実行例 `LRE.r` において, 乱数の初期値を変更しても少数の法則が成立することを確認せよ. また, パラメータ n, p の変更によって結果がどのように変わるか観察せよ.