極限定理

第 7 講 - 大数の法則・中心極限定理・少数の法則

村田 昇

講義概要

- 独立な確率変数の性質
- 大数の法則
- 中心極限定理
- 少数の法則

基本事項の確認

確率変数

- 乱数の数学モデル : 値がランダムに決定される変数
- 任意の区間 $[a, b]$ に含まれる確率が定められている
  - **数学的には厳密性を欠くが，本講義ではこの定義**
- 確率変数 X が区間 $[a, b]$ $(a \leq b)$ に含まれる確率

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

(特に $a = b$ のとき $P(X = a)$ と書く)

- 今回は有限個の値のみをとる確率変数を考える
  - **無限個の値，特に連続的な値については次回以降**

平均と分散

- 確率変数 X の観測値 : $x_1, x_2, \ldots, x_N$
- **平均** もしくは **期待値**

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

- **分散** (= 標準偏差^2)

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

例題

- 偏ったサイコロの問題

  確率変数 X は偶数の出る確率が奇数の 2 倍のサイコロの目を表すとする.

  $$P(X=1)=P(X=3)=P(X=5)=1/9$$
  $$P(X=2)=P(X=4)=P(X=6)=2/9$$

  このとき X の平均と分散を求めよ.

- 解答 (計算例)

  X の平均は

  $$\mathbb{E}[X]=\sum_{x=1}^{6} xP(X=x)=11/3=3.6666\ldots$$

  X の分散は

  $$\mathbb{E}[X^2]=\sum_{x=1}^{6} x^2P(X=x)=49/3$$
  $$\mathrm{Var}(X)=49/3-121/9=26/9=2.88\ldots$$

- 解答 (R を用いた計算例)

```r
#' 平均と分散の計算
p <- rep(c(1/9,2/9),3) # 確率の値 (1/9 と 2/9 を交互に 3 回繰り返す)
x <- 1:6 # サイコロの目の値
(mu <- sum(x*p)) # 平均値の計算
(v <- sum((x-mu)^2*p)) # 分散の計算
sqrt(v) # 標準偏差

#' 正規化しないで計算する方法もある
w <- rep(1:2,3) # 1,2 の繰り返し (確率ではない)
weighted.mean(x,w)
weighted.mean(x^2,w)-weighted.mean(x,w)^2
```

```
[1] 3.666667
[1] 2.888889
[1] 1.699673
[1] 3.666667
[1] 2.888889
```

独立性と同分布性

同時分布

- 観測データは確率変数の集合
- **確率変数列** $X_1, X_2, \ldots, X_n$ に対する考察が重要
- 定義

  "X_1 が x_1 という値をとり, X_2 が x_2 という値をとり, ..., X_n が x_n という値をとる" という事象が起きる確率を**同時分布**という.

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

独立性

- 無関係にサンプリングされた観測データの性質
- 定義

  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．

同分布性

- 同一の法則に従って生成された観測データの性質
- 定義

  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．

独立同分布性

- 一般に分析対象のデータには**独立性** と **同分布性**が同時に仮定される
- 定義

  独立かつ同分布である確率変数列を**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

無限列の独立性と同分布性

- 無限列に対しては任意の部分列について考える
- 独立性

  $X_1, X_2, \ldots$ 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．

- 同分布性

  $X_1, X_2, \ldots$ 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．

- 独立同分布性

  $X_1, X_2, \ldots$ 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

大数の法則の概要

- 要点

  同一の法則に従って生成された集団から**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** に近づく

- 例
  – 歪みの無いコインの表が出た回数の割合
  – 視聴率の調査
- この法則を数学的に定式化した定理が **大数の法則**

大数の強法則

- 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 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．

  これを “$\bar{X}_n$ は $n \to \infty$ のとき μ に **概収束** する” という．

実習

数値実験の設計

- 方針

  真の平均と標本平均を比較する．

  標本平均は観測データに依存するので，統計的な性質を見るには繰り返し実験 (Monte-Carlo 法) を行う．

  - 適当な分布を設定する (例 : 偏りのあるサイコロ)
  - n 個の確率変数 (乱数) の標本平均を計算する
  - 真の平均と標本平均の差を計算する
  - n を大きくしたときの差の性質を観察する

練習問題

- 大数の法則の数値実験を行いなさい．
  - 歪んだサイコロを例として，n 回サイコロを振って標本平均 (期待値) を求めたとき，n の値に応じて真の値と標本平均がどのくらい異なるか調べなさい．
  - n の値ごとに多数回実験を行い，標本平均の分布が n の値とともにどのように変化するか調べなさい．

中心極限定理

中心極限定理の概要

- 大数の法則の主張
  - n を大きくすると標本平均 $\bar{X}_n$ は真の平均 μ に近づく
  - 推定誤差 $\bar{X}_n - \mu$ は n を大きくすると 0 に近づく
  - どの程度の大きさになるのか定量的な評価は与えていない
- 誤差の評価の定量化とは
  - 推定誤差がある区間 $[\alpha, \beta]$ に入る確率で定量的に評価可能

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

  - 上式の正確な計算は一般には困難
- サンプル数が大きい場合の定量的な評価の近似方法を述べたのが **中心極限定理**

中心極限定理

- 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

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

  が成り立つ．

中心極限定理の意味

- X_i の分布が何であっても，サンプル数 n が十分大きければ，標本平均と真の平均の差 $\bar{X}_n - \mu$ の分布は**標準正規分布**で近似できる

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

  - 被積分関数 $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ を**標準正規密度**という

実習

数値実験の設計

- 方針

  規格化した標本平均と真の平均の差

  $$Z = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

  の分布と標準正規分布を比較する．

  - 中心極限定理が正しければ，十分小さいビン $[a, b]$ におけるヒストグラムの高さ (密度) は $\phi(a)$ で近似される
  - Z を多数観測し分布 (ヒストグラム) を求める
  - Z の分布と標準正規密度 $\phi(x)$ を比較する
    - 密度表示は `hist()` で `freq=FALSE` を指定
    - 標準正規密度 $\phi(x)$ は関数 `dnorm()` で計算可

練習問題

- 中心極限定理の数値実験を行いなさい．
  - 歪んだサイコロを例として，n 回サイコロを振って標本平均 (期待値) を求めたとき，n が大きければ正規化した値は標準正規分布に従うことを確認しなさい．
  - 確率 (歪み具合) が異なっても，上記の性質は変わらないことを確認しなさい．

少数の法則

少数の法則の概要

- 滅多に起きない事が起こる回数に関する法則
- 例：不良品発生率の低い工場での日々の不良品の個数の分布

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.

個数 S_n は確率変数であり，強度 λ の **Poisson 分布** で近似できる.

- この状況を正確に述べたのが **少数の法則**

少数の法則

- 定理の問題設定

  $X_1, X_2, \ldots, X_n$ 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

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

- 定理

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

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

  ならば，任意の整数 $k \geq 0$ に対して以下が成り立つ:

  $$P\Big(\sum_{i=1}^{n} X_i = k\Big) \to e^{-\lambda}\frac{\lambda^k}{k!} \quad (n \to \infty).$$

  - 定理の $\sum_{i=1}^{n} X_i$ が不良品の例の S_n に対応

Poisson 分布

- 定義

  確率変数 X の取りうる値が 0 以上の整数全体で，値が整数 $k \geq 0$ となる確率が

  $$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

  で与えられるものを強度 λ の **Poisson 型確率変数**その確率法則を強度 λ の **Poisson 分布** と呼ぶ.

実習

数値実験の設計

- 方針

  小さな確率で $X = 1$ となる確率変数を多数観測し，その合計値の分布を調べ，Poisson 分布と比較する.

  - 確率 $P(X = 1) = p$ を小さな値に設定する
  - n 個 (非常に多数) の確率変数の合計 S_n を計算する
  - S_n を多数観測し分布を求める
  - S_n の分布を強度 $\lambda = pn$ の Poisson 分布と比較する
    - 確率 p サイズ n の二項乱数 `rbinom()` が利用可能
    - Poissson 分布の確率値は関数 `dpois()` で計算可能

練習問題

- 少数の法則の数値実験を行いなさい.
  - 1 日の総生産量 (n) が 5000，不良品の発生確率 (p) が 0.002 である工場を例として，2 年間の操業 (週 5 日 x50 週間) において観測される不良品数の分布を確認しなさい.
  - 母数 n, p の違いによって結果がどのように変わるか観察しなさい.

補遺

重複対数の法則

- 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

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

  が成り立つ.
  - 大数の法則と中心極限定理の中間的な評価と考えることができる

Hartman-Wintner の定理

- 定理

  前定理の条件のもと，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.

次回の予定

- 一般の確率変数
- 離散分布
  - 離散一様分布・二項分布
  - Poisson 分布・幾何分布
- 連続分布
  - 一様分布・正規分布
  - ガンマ分布・t-分布・F-分布