データ解析基礎

第4回および第5回 確率モデル

確率モデルとは何か？

- 仮想的な法則
  - 現象がある法則に従って生じていると仮定する．
  - 法則を仮定したときに，どのようなデータが得られるかを予測する．
- なぜ確率モデルを考えるのか？
  - 「仮説」の表現
    法則性を仮定して結果を予測する．

母集団，標本，仮説の相互関係をよく理解しよう

代表的な確率モデル

3

- 離散値の分布
  - 二項分布：ベルヌーイ試行での成功回数
  - 幾何分布：ベルヌーイ試行で成功までの試行数
  - ポアソン分布：一定時間に時々起きる現象の回数
- 連続値の分布
  - 一様分布：確率密度関数が一定．
  - 指数分布：故障までの時間の分布．記憶の欠如．
  - 正規分布：連続値の分布で最も一般的なもの．
    普遍的に見られる．理論的取扱いが楽．

確率モデルの期待値と分散

4

- データに基づく平均と分散の定義（第１回）
  データとモデル，相対度数と確率を混同するな．
- 確率モデルの期待値と分散の定義
  確率変数 X が値 Xi をとる確率を$p(Xi)$としたとき

  期待値　$E(X) = \Sigma\, Xi\, p(Xi)$

  分散　$V(X) = \Sigma\, (Xi - E(X))^2 p(Xi) = E\,((X - E(X))^2)$

  $= E\,(X^2) - (E(X))^2$

  連続値の分布の場合は積分をとる（後述）．

ベルヌーイ試行 (Bernoulli sequence)

5

- 各試行において場合の数が二つしかない．
- 各試行では同じことを繰り返す．
- 各試行は独立である．
- 例：
  - コインを投げて表裏のどちらがでるか？
  - さいころの目の和が奇数か偶数か？
  - 試験に合格するかどうか？
  - ビールＡとＢのどちらを好むか？
  - 刺激が見えるか見えないか？

二項分布 (binomial distribution)

6

- n 回のベルヌーイ試行において m 回成功する確率を表す分布．
- 関係する値
  - 全体の試行数 n
  - 個々の試行において成功する確率 p
  - 成功する回数 m：これが確率変数

    n と p を固定して，m が0から n まで変えたときの確率分布

二項分布の導出(1)

7

- 一つ一つ計算した場合（例：$n = 3$）
- 成功の確率を p, 失敗の確率をqとする．
  - 失敗は成功の補事象だから $q = 1 - p$
- $m = 0, 1, 2, 3$の場合をそれぞれ計算する．
  - 事象の独立性と排反性に注意する．
    事象が排反だから，確率の和が意味をもつ．
    事象が独立だから，確率の積が意味をもつ．
- 得られた数値の列が確率分布である．
  cf. 確率分布と相対度数分布を混同するな．

二項分布の導出(2)

8

- 系統的に考えた場合
  - n 回の試行において，m 回成功する場合の数は？
- n 個の異なる球からm 個を取出すときの組合せと同じ
  - したがって，$p(X = m) = nCm\, p^m q^{n-m}$
    これが，2項分布の確率を与える式である．
- すべてのmについて和をとると 1 になることを確認．

二項分布の形

9

- n, m, pを変えて分布の形がどうなるかを観察してみる．
  - p の値によって，分布の形はどう変わるか？
  - n が大きくなると，分布の形はどうなるか？
- MATLABのデモプログラムで見てみる．

二項分布の期待値と分散

10

- 期待値：

  $E(X) = np$

- 分散：

  $V(X) = npq$

  - こういった式の導出をうのみにしていると，あとで訳がわからなくなる．
  - 定義に従って自分の手で計算してみるのが大事．

    期待値は板書で．分散については演習問題．

ベルヌーイ分布

11

- ベルヌーイ試行一回の成功／失敗を与える分布
  - 確率変数のとり得る値は0と1のみ．

    $p(X = 0) = 1 - p$

    $p(X = 1) = p$
- すぐにわかるように，確率の総和は1になる．
- 期待値と分散は？
  - 直感的に理解しにくいが，定義より計算できる．
    期待値は板書で．分散は演習問題．

ベルヌーイ分布と二項分布の関係

12

- 二項分布は
  - ベルヌーイ分布に従う独立な確率変数の和の分布である．
- 同一分布に従う独立な確率変数の和の期待値と分散
  - 事象が独立であれば，積事象の確率は確率の積．
    独立性の性質を思い出せ！
  - このことを使うと，和の期待値と分散が計算できる．

ベルヌーイ分布と二項分布の関係

13

- 結果は

$$E(X1 + X2 + \ldots + Xn) = n\,E(X)$$

$$V(X1 + X2 + \ldots + Xn) = n\,V(X)$$

  - 分散の証明には，独立な変数の共分散が0であることを使う．

- ベルヌーイ分布と二項分布で以上の関係を確かめよ．

幾何分布

14

- ベルヌーイ試行においてn回目で初めて成功する確率
  - n-1回目まですべて失敗する確率はq^{n-1}
  - n回目に成功する確率はp
- 独立性より，求める確率は

$$P(X = n) = pq^{n-1}$$

- 確率変数の取りうる値は1から無限大までになる．
- 分布の形を見てみよう．

幾何分布の性質

15

- 期待値と分散

  期待値：$E(X) = 1 / p$

  分散：　$V(X) = q^2 / p$

- 性質：記憶の欠如

  1. 試行を始める前の時点で，n 回目に成功する確率．

  2. k 回未成功のとき，$(k+n)$回目に成功する確率．

     人情としては2.の方が確率が高いように思えるが，
     実際は等しい．「自然は記憶を持たない」

ポアソン分布(Poisson distribution)

16

- 単位時間にある試行が何回生じるかを表わす分布

  - たまにしか生じないことがある時間に起きる確率

    原子核崩壊，交通事故，飛行機事故

    客の来店，品物の発注，電話の発信

- 二項分布の期待値$(m = np)$を保ち，$n \to \infty$の極限．

  - p が非常に小さいベルヌーイ試行で膨大な試行を繰り返したときに k 回成功する確率

$$p(X = k) = nCk \quad p^k \quad q^{n-k}$$

$$n \to \infty, \quad p \to 0$$

ポアソン分布の性質

17

- ポアソン分布の形

$$P(X = k) = \frac{\exp(-m)\, m^k}{k!}$$

- 期待値と分散
  - 期待値： $E(X) = m$　二項分布の np に相当．
  - 分散： $V(X) = m$　npq に相当
- 確率和が1になるのは指数関数の展開式からわかる．
- 二項分布からポアソン分布を導く計算は省略．

連続値の確率分布

18

- 個々の点に対しては確率は決まらない（確率は0）．
  - 離散値の確率とは考え方が違う．
- 変数がある区間に入る確率は求められる．

$$P = P(x - \Delta x \leqq X \leqq x + \Delta x)$$

- 確率密度関数の考え方(probability density function)
  - この確率を区間幅 $2\,\Delta x$ で割り，$\Delta x \to 0$ とした極限

$$f(x) = \lim \frac{P(x - \Delta x \leqq X \leqq x + \Delta x)}{2\,\Delta x}$$

全区間の確率1を出発点にして区間を半分と割っていけばイメージがつかめる．

確率と確率密度関数，確率分布関数

19

- 確率は「確率密度関数の積分」で求められる．

  確率は面積（密度×区間幅）で決まる．

- 分布関数
  - 確率密度関数を無限小から積分したもの．
  - 累積相対度数分布に対応する．
  - 累積分布関数(cumulative density function)ともいう．

変数の変換による確率密度関数の変換

20

- 変数を変換すると確率密度関数はどう変わるか？
  - $Y = 2X$ なる変数Yを考えたとき $g(y)$ はどうなるか？
- 陥りやすい誤り： $g(y) = f(x)$
  - 確率が面積であることを忘れている．
  - 正解は：$g(y) = 0.5 \times f(x)$
- 一般には，

$$g(y) = \frac{1}{\frac{dy}{dx}} f(x)$$

この係数が区間の伸縮比を与えている．
区間が伸びた分密度を薄める必要がある．

一様分布 (uniform distribution)

21

- 変数の値によらず確率密度が等しい分布
- 区間を $x_1 \leqq x \leqq x_2$ とすれば，
  - 密度関数は，　$f(x) = 1 / (x_2 - x_1)$
  - 期待値と分散は，

$$E(X) = (x_1 + x_2) / 2$$ （板書で）

$$V(X) = (x_2 - x_1)^2 / 12$$ （各自確かめよ）

- 例えば，[0, 1]の一様乱数の期待値は1/2，分散は1/12．

指数分布 (exponential distribution)

22

- 機械の寿命(故障までの時間)を表す分布として使う．
- 確率密度関数は次式で与えられる（λは正の値）．

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geqq 0 \\ 0 & \text{otherwise} \end{cases}$$

- 期待値と分散：

$$E(X) = 1 / \lambda \qquad V(X) = 1 / \lambda^2$$

- 記憶の欠如：幾何分布と同様に記憶をもたない．
  - 累積分布関数を考えて，条件付き確率を計算する．
  - 故障率が寿命とともに変動する場合は×．

正規分布 (normal distribution)

- 確率密度関数
  - 二つのパラメータ μ と σ で指定される．

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

  - 期待値と分散は，それぞれ μ と σ^2 になる．
- 通常，$N(\mu, \sigma^2)$ と表記される．
- $\mu = 0$，$\sigma = 1$ の分布を「標準正規分布」といい，$N(0, 1)$ と表す．

正規分布の性質

- 世の中のほとんどの現象を近似する分布
- 平均値に関して対称で，左右に無限の広がりをもつ．
- 期待値と標準偏差（分散）の二つのパラメータで決定される．
- 正規分布に従う独立な変数の和も正規分布に従う．
  - これを再生性という．
  - 二項分布，ポアソン分布も再生性をもつ．

正規分布の歪度と尖度

25

- 歪度と尖度の意味（復習）
  - 歪度：確率密度関数の対称性を表す指標
  - 尖度：確率密度関数の集中性を表す指標
- 歪度と尖度の定量的定義
  - 歪度，尖度はそれぞれ3次，4次のキュミュラントにあたる．
  - キュミュラントとは，積率母関数の対数をTaylor展開したときの係数（詳細は省略）．
  - 正規分布では 3次以上のキュミュラントがすべて0．
    したがって，歪度，尖度ともに0となる．

チェビシェフ (Chevyshev)の不等式

26

- 「ほとんどのデータは平均の周りにあること」を示した定理．
- 数式でかけば，

$$\mathrm{Prob}(\,|X-\mu| \geqq k\sigma\,) \leqq 1/k^2$$

  平均から標準偏差のk 倍遠ざかると，その外側の区間にXがある確率は $1/k^2$ の速さで減っていく．

- 分布の形によらずに成立するところがミソ．
- 証明は単純だが省略（教科書を参照のこと）．

大数の法則 (law of large numbers)

27

- $X1, X2, \ldots, Xn$が 独立に同じ分布（期待値 μ,分散σ^2 ）に従っているとする．このとき，それらの平均$\overline{X}^{(n)}$について，次の二つの法則が成立する．
- 大数の弱法則
  - $n \rightarrow$ 無限大で，$P(|\overline{X}^{(n)} - \mu| > \varepsilon)$ はいくらでも 0に近くなる．チェビシェフの不等式から導ける．
    （$\overline{X}^{(n)}$ の期待値，分散はそれぞれどうなるか？）
- 大数の強法則
  - $n \rightarrow$ 無限大で，$P(|\overline{X}^{(n)} - \mu| < \varepsilon)$ はいくらでも1に近くなる．こちらは証明が難しい．

中心極限定理 (central limit theorem)

28

- $X1, X2, \ldots, Xn$が独立に同じ分布（期待値μ, 分散σ^2)に従っているとする．このとき，それらの平均 $\overline{X}^{(n)}$ の分布は，もとの分布にかかわらず，n が大きくなるにしたがって正規分布に近づく．
- 正確には，平均 $\overline{X}^{(n)}$からμを引き, $(\sigma / \sqrt{n})$で割った値

$$\frac{\overline{X}^{(n)} - \mu}{\sigma / \sqrt{n}}$$

は，n が大きくなると標準正規分布に近づく．

証明は省略（教科書を参照のこと）．

中心極限定理の例

29

- 離散分布の例
  - 二項分布は独立なベルヌーイ試行の和の分布．
    → 二項分布でnを大きくすれば正規分布に近づく．
    （同様にしてポアソン分布も正規分布に近づく）
- 連続分布の例
  - 一様分布に従う独立変数の和は正規分布に従う．
    → 正規分布に従う乱数生成法の一つとして，
    「一様乱数を12個足して6を引く」がある．

中心極限定理の意味：正規分布の重要性

30

- どのような確率変数であっても，多数集めてその平均をとれば，その平均値は正規分布に従う．
- データを多数集めると個々のデータの意味が薄れる．
  - 個々のデータが従う分布は見えなくなってしまう．
- 多数回の実験により得られた平均値は，正規分布に従うものと考えてよい．
  - 次回以降議論する検定や推定は，得られたデータが正規分布に従っていることを前提にしている．