応用システム工学
第3回　モデリング

平成22年6月11日

指数型分布族

なぜ指数型分布族が良いのかは，一般化線形モデルで述べる

- 確率分布が指数関数で表される

$$f(x;\theta) = s(x)\tau(\theta)e^{a(x)b(\theta)}$$

  – 二項分布　→　サイコロの目の出る確率

$$P(x) = {}_nC_x p^x (1-p)^{n-x}$$

  – ポアソン分布　→　一定の期間において，ごくまれに起こる事象の確率分布

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

  – 正規分布　→　自然界で起こる現象の多くがその分布に当てはまる(特に期待値に関する分布)
    - 二項分布，ポアソン分布の正規分布による近似もある

指数型分布族としてのポアソン分布

- ポアソン分布→離散事象の確率分布
- 確率変数　$x_1, \cdots, x_n$
  - 値域　$x_i = 0,1,2,\cdots$
  - パラメータ θ　$f(x_i;\theta) = \dfrac{\theta^{x_i} e^{-\theta}}{x_i!}$

$$= \exp\left(x_i \log\theta - \theta - \log x_i!\right)$$

- ポアソン分布の同時分布

$$f(x_1,\cdots,x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

$$= \frac{\theta^{x_1} e^{-\theta}}{x_1!} \times \frac{\theta^{x_2} e^{-\theta}}{x_2!} \times \cdots \times \frac{\theta^{x_n} e^{-\theta}}{x_n!} = \frac{\theta^{\sum x_i} e^{-n\theta}}{x_1! x_2! \cdots x_n!}$$

指数型分布族としての二項分布

- 二項分布→成功・失敗やさいころの目
- 確率変数Y→n回の独立試行における成功回数
  - 確率関数
  - y=0,1,2,・・・,n
  - 成功の確率π
    - 指数表示

$$f(x;\pi)=\binom{n}{x}\pi^{x}(1-\pi)^{n-x}$$

$$X \sim binominal(n,x)$$

$$f(x;\pi)=\exp\left[x\log\pi - x\log(1-\pi)+n\log(1-\pi)+\log\binom{n}{x}\right]$$

正規分布から導かれる分布

- 推定量，検定統計量の標本分布の多くは正規分布と関係
  - 正規分布に従う確率変数から導かれる場合
  - 大標本に対する中心極限定理より漸近的に関係する場合
    - 正規分布　→　正規分布そのもの
    - カイ二乗分布(自由度n)　→　標準正規分布に従うn個の独立な確率変数の二乗和の分布
    - t分布(自由度n)　→　2つの独立な確率変数の比の分布。分子は標準正規分布に従う確率変数，分母は中心カイ二乗分布に従う確率変数を自由度で除したものの平方根
    - F分布　→　2つの独立したカイ二乗確率変数(自由度n,m)を各々の自由度で除したものの比
      - 中心F分布　→　中心カイ二乗分布のみを用いる
      - 非心F分布　→　分子が非心カイ二乗分布，分母が中心カイ二乗分布

中心極限定理(central limit theorem)
ある弱い条件さえ満たせば、どんな確率分布を持つ確率変数の和でも、同じ釣鐘型曲線（正規分布）に近づく。証明Lyapunov(1901)

t分布

- t分布(自由度n)
  - 二つの独立な確率変数の比
    - 標準正規分布に従う確率変数 $Z \sim N(0,1)$
    - 中心カイ二乗に従う確率変数 $X^2 \sim x^2(n)$
      - 自由度nで割って平方根で割ったもの

$$T = \frac{Z}{\sqrt{\frac{X^2}{n}}} \sim t(n)$$

ただし，ZとX2は独立

中心F分布

- 独立な中心カイ二乗確率変数
  - X_1^2:自由度n　$X_1{}^2 \sim x^2(n)$
  - X_2^2:自由度m　$X_2{}^2 \sim x^2(m)$
- 確率変数各々の自由度割ったものの比

$$F = \frac{\dfrac{X_1{}^2}{n}}{\dfrac{X_2{}^2}{m}} \sim F(n, m)$$

中心F分布とt分布の関係

- t分布の二乗

$$T^2 = \left(\frac{Z}{\sqrt{\frac{X^2}{n}}} \right)^2 = \frac{\frac{Z^2}{1}}{\frac{X^2}{n}} \sim F(1, n)$$

- 自由度nのt分布t(n)に従う確率変数の二乗は自由度(1,n)のF分布F(1,n)に従う

非心F分布

- 二つの独立な確率変数を各々の自由度で割ったものの比
  - 分子　非心カイ二乗分布の確率変数 $X_1{}^2 \sim x^2(n, \lambda)$
  - 分母　中心カイ二乗分布の確率変数 $X_2{}^2 \sim x^2(m)$

$$F = \frac{\frac{X_1{}^2}{n}}{\frac{X_2{}^2}{m}}$$

  - 非心F分布の平均は，同じ自由度の中心F分布の平均よりも大きい

統計モデルの構築
モデリングのプロセス

- モデルの特定
  - 反応変数と説明変数の関係式
  - 反応変数の確率分布

  を決める

- モデルパラメータの同定
  - 最小二乗法, 最尤推定法
- モデルの検証
  - モデルの仮説検定
- モデルを用いた検討
  - 信頼度区間

統計モデル構築のためのデータ解析

- 統計的なデータ解析
  - データの検討→異常値の除去，モデルの定式化
  - 解析項目
    - 測定尺度
      - 量的尺度
      - 質的尺度
        - » 名義尺度(赤青黄色，はい･いいえ)
        - » 順序尺度(大中小，年齢)
    - 確率分布→分布性状の決定
    - 変数の関連性
      - 質的な変数間
      - 連続変数間
      - 質的な変数と連続変数間

モデルの定式化
(一般化線形モデル)

- モデルの構成
  - 要素
    - 確率変数
    - 説明変数(複数)
  - 確率分布
    - 一般化線形モデル←指数型分布族の確率分布
      - 正規分布，二項分布，ポアソン分布
  - 期待値と説明変数の関係式→線形結合

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

線形モデル

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- 反応変数Y1,…,YN全体に対する行列表記

$$g[E(y)] = X\beta$$

  - ただし

反応ベクトル

$$\underline{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}$$

関数のベクトル
(gは共通)

$$g[E(y)] = \begin{bmatrix} g[E(Y_1)] \\ \vdots \\ g[E(Y_N)] \end{bmatrix}$$

パラメータ
ベクトル

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}$$

  - X:説明変数の水準もしくは測定値
    - 量的な説明変数
    - 質的な説明変数　含む・含まない:ダミー変数
      (0,1の時は指示変数)

パラメータの同定

- 推定法
  - 最小二乗推定法
    - 期待値の式(分散共分散の式)が必要
    - 確率変数Yiの分布は不要
  - 最尤推定法
    - 確率変数Yiの同時確率分布が必要
  - 一般的に最尤推定量と最小二乗推定量は同じ

最小二乗推定1

- 独立なn個の確率変数$Y_1, Y_2, \cdots, Y_n$
  - 期待値 $\mu_1, \mu_2, \cdots, \mu_n$
    - 仮定
      - 期待値はパラメータベクトルβの関数として表される

$$E(Y_i) = \mu_i(\beta) \qquad \beta = [\beta_1, \beta_2, \cdots, \beta_p]^t, p < n$$

- 最小二乗法
  - 確率変数(観測値)Y_iと期待値μ_iの差の平方和を最小にする推定量$\hat{\beta}$を見つける
    - 評価関数 $S = \sum [Y_i - \mu_i(\beta)]^2$

最小二乗推定2

- 推定量 $\hat{\beta}$ の最適値→評価関数が最小
  - 評価関数Sを各要素 β_j で微分
    - 連立方程式の解

$$\frac{\partial S}{\partial \beta_j} = 0,\ j = 1,2,\cdots,p$$

- 確率変数 Y_i 間の分散 σ_i^2 が等しくない場合
  - 分散の大きい(信頼性が低い)確率変数の影響を抑える
    - 重み付き平方和を最小にする

$$S = \sum w_i \left[Y_i - \mu_i(\beta) \right]^2$$

重み $w_i = \dfrac{1}{\sigma_i^{\ 2}}$

最小二乗推定3

- 重み付き最小二乗推定量の一般表現
  - 確率変数ベクトル $y = \left[Y_1, Y_2, \cdots, Y_n\right]^t$
  - 平均(期待値)ベクトル $\mu = \left[\mu_1, \mu_2, \cdots, \mu_n\right]^t$
  - 分散共分散行列*V*
    - 行列の対角要素σ_i^2(分散)
    - 行列の非対角要素$\rho_{ij}\sigma_i\sigma_j$(共分散)
      - ρ_{ij}:Y_iとY_jの相関係数
  - 評価関数

$$S = (Y - \mu)^t V^{-1} (Y - \mu)$$