応用システム工学
第二回　確率統計の基礎

平成22年06月04日
正規分布の分散
多変数の確率分布
一般化線形モデル

指数型分布族

- 確率分布が指数関数で表される

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

  - 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

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

期待値と分散

- 期待値の線形性
  - 前回正規分布関数の期待値を求めた

$$E[A+B]=E[A]+E[B]$$

- 期待値の線形性を用いた分散σ^2の性質

$$\begin{aligned}\sigma^2 &= V[X]=E\left[(X-E[X])^2\right]\\ &=E\left[X^2-2XE[X]+E[X]^2\right]\\ &=E\left[X^2\right]-2E[X]E[X]+E[X]^2\\ &=E\left[X^2\right]-E[X]^2\end{aligned}$$

正規分布関数の分散1

- 正規分布の確率密度関数(確率変数X)

$$f_X\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- 分散

$$V[X]=\int_{-\infty}^{\infty}(x-\mu)^2\,f_X(x)dx$$

$$=\int_{-\infty}^{\infty}\left(x^2-2x\mu+\mu^2\right)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$

$$=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\left(x^2-2x\mu+\mu^2\right)\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$

正規分布関数の分散2

$$V[X]=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}x^2\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$
$$-\frac{2\mu}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}x\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$
$$+\frac{\mu^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$

正規分布関数の分散3

- 第3項 正規分布関数の確率密度関数1参照

$$\frac{\mu^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx=\mu^2\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$

$$=\mu^2$$

- 第2項 正規分布関数の期待値2参照

$$\frac{-2\mu}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}y\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx=\frac{-2\mu}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}y\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$

$$=-2\mu^2$$

正規分布関数の分散4

- 第1項
  - 変数変換

$$z = \frac{x-\mu}{\sigma} \quad \Rightarrow \quad \frac{dz}{dx} = \frac{d}{dx}\frac{x-\mu}{\sigma} = \frac{1}{\sigma}$$

$$x = \sigma z + \mu \qquad x : -\infty \leftrightarrow \infty \Leftrightarrow z =: -\infty \leftrightarrow \infty$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} x^2 \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\left(z\sigma+\mu\right)^2 e^{-\frac{1}{2}z^2}\sigma dz$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\left(z^2\sigma^2+2z\sigma\mu+\mu^2\right)e^{-\frac{1}{2}z^2}\sigma dz$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(z^2\sigma^2+2z\sigma\mu+\mu^2\right)e^{-\frac{1}{2}z^2}dz$$

正規分布関数の分散5

$$\int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2}\, dz = 0 \qquad \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}z^2\right] dz = \sqrt{2\pi}$$

$$\int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2}\, dz = \int_{-\infty}^{\infty} z z e^{-\frac{1}{2}z^2}\, dz = \left[z\left(-e^{-\frac{1}{2}z^2} \right) \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-\frac{1}{2}z^2}\, dz$$

$$= \left[0 - 0\right] + \sqrt{2\pi} = \sqrt{2\pi}$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(z^2\sigma^2 + 2z\sigma\mu + \mu^2 \right) e^{-\frac{1}{2}z^2}\, dz = \frac{1}{\sqrt{2\pi}} \left(\sigma^2\sqrt{2\pi} + 2\sigma\mu \times 0 + \mu^2\sqrt{2\pi} \right)$$

$$= \sigma^2 + \mu^2$$

正規分布関数の分散6

$$V[X]=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}x^2\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$

$$-\frac{2\mu}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}x\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$

$$+\frac{\mu^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx$$

$$=\left\{\sigma^2+\mu^2\right\}+\mu^2-2\mu^2=\sigma^2$$

多変量の確率分布1
2変数

- 同時確率密度関数
  - 二つの確率変数X,Y

$$x \le X \le x + dx$$
$$y \le Y \le y + dy$$

かつ
となる確率が

$$f_{XY}(x, y)dxdy$$

と書ける

  - XとYの同時確率密度関数

$$f_{XY}(x, y)$$

多変数の確率分布2
n変数

- 同時確率密度関数
  - n個の確率変数

$$X_1, X_2, \cdots, X_n$$

  - 各々 $x_i \le X_i \le x_i + dx_i \left(1 \le i \le n\right)$ となる確率が

$$f_{X_1 X_2 \cdots X_n}\left(x_1, x_2, \cdots, x_n\right) dx_1 dx_2 \cdots dx_n$$

  - $X_1, X_2, \cdots, X_n$ の同時確率密度関数

$$f_{X_1 X_2 \cdots X_n}\left(x_1, x_2, \cdots, x_n\right)$$

多変数の確率分布3
2変数

- 周辺密度関数
  - 二つの確率変数X,Y
  - Yを無視したXのみに関する確率密度関数を指す

$$f_X(x)$$

  - 同時確率密度関数 $f_{XY}(x,y)$ と周辺密度関数 $f_X(x)$ の関係

$$f_X(x)=\int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

多変数の確率分布4
n変数

- 周辺密度関数
  - n個の確率変数

$$X_1, X_2, \cdots, X_n$$

    - m個のみに注目した同時確率密度

$$f_{X_1 X_2 \cdots X_m}\left(x_1, x_2, \cdots, x_m\right)$$

    - $X_1, X_2, \cdots, X_m$ についての周辺密度関数

$$f_{X_1 \cdots X_m}\left(x_1, \cdots, x_m\right) = \int_{-\infty}^{\infty} dx_{m+1} \cdots \int_{-\infty}^{\infty} dx_n \, f_{X_1 \cdots X_n}\left(x_1, \cdots, x_n\right)$$

確率変数の独立性1

- 二つの確率変数X,Y
- 独立性の定義

$$\Leftrightarrow P\left(X \in A \text{かつ} Y \in B\right) = P\left(X \in A\right)P\left(Y \in B\right)$$

$$\Leftrightarrow f_{XY}\left(x, y\right) = f_X\left(x\right)f_Y\left(y\right)$$

確率変数の独立性2

- n個の確率変数

$$X_1, X_2, \cdots, X_n$$

- 独立性の定義

$$\Leftrightarrow P\left(X_1 \in A_1 \text{かつ} X_2 \in A_2 \text{かつ} \cdots \text{かつ} X_n \in A_n\right)$$

$$= P\left(X_1 \in A_1\right)P\left(X_2 \in A_2\right)\cdots P\left(X_n \in A_n\right)$$

$$\Leftrightarrow f_{X_1 X_2 \cdots X_n}\left(x_1, x_2, \cdots, x_n\right) = f_{X_1}\left(x_1\right) f_{X_2}\left(x_2\right)\cdots f_{X_n}\left(x_n\right)$$

- 自由度
  - 独立に選べる確率変数の数

条件付き確率1

- 二つの確率変数X,Yが必ずしも独立でない

$$f_{XY}(x,y) = f(y \mid x) f_X(x)$$

  - YのXに関する条件付き確率密度　$f(y \mid x)$
    - Xが値xを取る条件下でYが値yを取る確率密度
      - XとYが独立な場合　$f(y \mid x) = f_Y(y)$
- 同時確率密度$f_{XY}(x,y)$と条件付確率密度$f(y \mid x)$の関係

$$f(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int_{-\infty}^{\infty} f_{XY}(x,y)dy}$$

条件付き確率2

- 条件　$X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m$
  に対する，$X_{m+1} = x_{m+1}, X_{m+2} = x_{m+2}, \cdots, X_n = x_n$
  となる確率密度関数

$$f\left(x_{m+1}, x_{m+2}, \cdots, x_n \mid x_1, x_2, \cdots, x_m\right)$$

$$f\left(x_1, x_2, \cdots, x_m, x_{m+1}, \cdots, x_n\right)$$
$$= f\left(x_{m+1}, x_{m+2}, \cdots, x_n \mid x_1, x_2, \cdots, x_m\right)$$
$$\times f\left(x_1, x_2, \cdots, x_m\right)$$

確率変数の和の分布

- 二つの独立な確率変数X,Y
- Z=X+Yの分布
  - 事象　$Z = z$

$$\Leftrightarrow X + Y = z$$

$$\Leftrightarrow Y = z - x \text{かつ} X = x$$

X,Yが独立の時，Y=z-xかつX=xとなる確率

$$P_X(x)P_Y(z-x)$$

  - 離散分布　$X \in M$
  - Z=zとなる確率

$$P(Z = z) = \sum_{x \in M} P_X(x)P_Y(z-x)$$

用語(最初にするべき)

- 測定値または観測値
  - 説明変数
  - 予測変数
  - 独立変数
- 確率変数(上述の変数に応じて変動)
  - 反応変数
  - 結果変数
  - 従属変数

説明変数の種類

- 一般化線形モデル
  - 反応変数Yと説明変数$x_1, x_2, \cdots, x_m$の関係

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

    - 量的な説明変数x
      - パラメータβは，説明変数の変化に伴う反応変数の変化を表す
    - 質的な説明変数
      - 反応変数にパラメータが含まれるか否かを表す
      - ダミー変数，(0,1)の場合は指示変数

相関係数

- 二つの確率変数の間の類似の度合いを示す統計学的な指標
  - 二組の数値からなるデータ列

$$(x, y) = \{(x_i, y_i)\}, i = 1, \cdots, n$$

  - 相関係数 ρ_{ij}

$$\rho_{ij} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$\bar{x}, \bar{y}$ は各々の相加平均

共分散

- 共分散
  - 二組の対応するデータ間での，平均からの偏差の積の平均値

多変量正規分布

- 正規分布の確率変数 $Y_i \sim N\left(\mu_i, {\sigma_i}^2\right), i = 1, \cdots, n$
- 確率変数Yi,Yj間の共分散

$$\mathrm{cov}\left(Y_i, Y_j\right) = \rho_{ij}\sigma_i\sigma_j \qquad \rho_{ij} : Y_i \text{と} Y_j \text{の相関係数}$$

  - Y1,Y2,・・・,Ynの同時分布は多変量正規分布となる
    - 確率変数ベクトル $y = \left[Y_1, \cdots, Y_n\right]^t$
    - 平均ベクトル $\mu = \left[\mu_1, \cdots, \mu_n\right]^t$
    - 分散共分散行列V:対角要素σ_i^2,非対角要素$\rho_{ij}\sigma_i\sigma_j$

$$y \sim N\left(\mu, V\right)$$

正規分布の線形結合

- 正規分布の確率変数 $Y_i \sim N\left(\mu_i, \sigma_i^{\,2}\right), i = 1, \cdots, n$
- 確率変数の線形結合W

$$W = a_1 Y_1 + a_2 Y_2 \cdots + a_n Y_n$$

  - ただしa1,a2,･･･,a3は定数
  - Wは正規分布となる
    - 平均 $\sum_{i=1}^{n} a_i \mu_i$
    - 分散 $\sum_{i=1}^{n} a_i^{\,2} \sigma_i^{\,2}$

$$W = \sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^{\,2} \sigma_i^{\,2} \right)$$

カイ二乗分布1

- 自由度nの中心カイ二乗分布
  - 独立なn個の標準正規分布の確率変数 $Z_1, \cdots, Z_n$ の二乗和の分布

$$X^2 = \sum_{i=1}^{n} Z_i^2 = z^T z \sim x^2(n)$$

    - ただし $z = [Z_1, \cdots, Z_n]^t$ $\quad z^T z = \sum_{i=1}^{n} Z_i^2$
  - 期待値 $E(X^2) = n$
  - 分散 $\mathrm{var}(X^2) = 2n$

カイ二乗分布

- 独立なn個の正規分布の確率変数 $Y_1, \cdots, Y_n$ に対して $Y_i \sim N\left(\mu_i, {\sigma_i}^2\right)$
  - 平均 μ_i, 分散 ${\sigma_i}^2$
  - 標準正規分布化 $Z_i = \dfrac{Y_i - \mu_i}{\sigma_i} = N(0,1)$
  - カイ二乗分布

$$X^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2 \sim x^2(n)$$

非心カイ二乗分布

- 標準正規分布の確率変数 $Z_1, \cdots, Z_n$
- 正規分布の確率変数化 $Y_i = Z_i + \mu_i$
  - 確率変数の二乗和

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} (Z_i + \mu_i)^2 = \sum_{i=1}^{n} Z_i^2 + 2\sum_{i=1}^{n} Z_i \mu_i + \sum_{i=1}^{n} \mu_i^2 \sim x^2(n, \lambda)$$

- 自由度:n
- 非心パラメータ $\lambda = \sum_{i=1}^{n} \mu_i^2$
- 平均 $n + \lambda$
- 分散 $2n + 4\lambda$

カイ二乗分布

- 多変量正規分布の確率変数ベクトル

$$y = \left[Y_1, \cdots, Y_n\right]^t$$

$$y \sim N\left(\mu, V\right)$$

  - 分散共分散行列Vが正則，逆行列を持つ時

$$X^2 = \left(y-\mu\right)^T V^{-1}\left(y-\mu\right) \sim x^2\left(n\right)$$

  - 一般に多変量正規分布に対して
    確率変数 $y^T V^{-1} y$ は非心カイ二乗分布となる

$$x^2\left(n, \lambda\right) \qquad \lambda = \mu^T V^{-1} \mu$$

カイ二乗分布の再生性

- 中心・非心カイ二乗分布に従う独立な確率変数　$X_i^2 \sim x_i^2(n_i, \lambda_i)$　の和のカイ二乗分布
  - 自由度　$\sum_{i=1}^{n} n_i$
  - 非心パラメータ　$\sum_{i=1}^{n} \lambda_i$

$$\sum_{i=1}^{m} X_i^2 \sim x_i^2\left(\sum_{i=1}^{n} n_i, \sum_{i=1}^{n} \lambda_i\right) \Rightarrow$$ カイ二乗分布の再生性

カイ二乗分布の性質

- 多変量正規分布の確率変数(要素数n)

$$y \sim N(\mu, V)$$

  - 分散共分散行列Vの階数k<nで特異の場合
    - 逆行列が定まらない
    - Vの一般逆行列V-に対して
    - 確率変数 $y^T V^- y$ の性質→非心カイ二乗分布
      - 自由度k
      - 非心パラメータ $\lambda = \mu^T V^- \mu$

ポアソン分布2

- ポアソン分布に対する尤度関数　$L(\theta; y_1, \cdots, y_n)$
- 最尤推定値　$\hat{\theta}$
- 対数尤度関数

$$l(\theta; y_1, \cdots, y_n) = \log L(\theta; y_1, \cdots, y_n)$$
$$= (\sum y_i)\log\theta - n\theta - \sum(\log y_i!)$$

- 対数尤度関数の導関数

$$\frac{d}{d\theta} l(\theta; y_1, \cdots, y_n) = \frac{d}{d\theta}\left[(\sum y_i)\log\theta - n\theta - \sum(\log y_i!)\right]$$
$$= \frac{1}{\theta}\sum y_i - n$$

ポアソン分布3

- 最尤推定値が極値を取る条件

$$\frac{d}{d\theta} l\left(\hat{\theta}; y_1, \cdots, y_n\right) = \frac{1}{\hat{\theta}} \sum y_i - n = 0$$

$$\hat{\theta} = \frac{\sum y_i}{n} = \bar{y}$$

- 最大推定値の条件

$$\frac{d^2}{d\theta^2} l\left(\hat{\theta}; y_1, \cdots, y_n\right) = -\frac{1}{\theta^2} \sum y_i < 0$$

$\theta = \hat{\theta}$ で最大となる。最尤推定値は $\bar{y}$

二次形式の性質

- 二次形式とは
  - 全ての項の次数が2の多項式
    - ○ $y_1^2 + y_2^2$　　$2y_1^2 + y_2^2 + 3y_1y_2$
    - × $y_1^2 + y_2^2 + 2y_1$　　$y_1^2 + 3y_2^2 + 2$
- 行列の二次形式
  - 対称行列A　　$a_{ij} = a_{ji}$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

    - yiの二次形式
      - ○ $y^T Ay = \sum_i \sum_j a_{ij} y_i y_j$
      - ×(yi-μiの二次形式)

$$(y-\mu)^T V^{-1} (y-\mu)$$

正定値

- 二次形式 $y^T Ay$ と行列A
- 要素の全てが0でないy
- 正定値 $y^T Ay > 0$
  - 必要十分条件
    - 全ての行列式が正

$$\left|A_1\right| = a_{11} \quad \left|A_2\right| = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \left|A_3\right| = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

行列Aの階数は二次形式 $Q = y^T Ay$ の自由度 $\left|A_n\right| = \det A$

Cochranの定理

- 正規分布 $N\left(0,\sigma^2\right)$ の独立な確率変数 $Y_1,\cdots,Y_n$
- 二次形式 $$Q=\sum_{i=1}^{n}Y_i^2$$
  - 和 $$Q=Q_1+\cdots+Q_k$$
  - ただし $$Q_i\text{の自由度}m_i\left(i=1,\cdots,k\right)$$
    $$m_1+m_2+\cdots+m_k=n\text{の時に限り}$$
    $$Q_i,\cdots,Q_k\text{は独立な確率変数}$$
    $$\frac{Q_1}{\sigma^2}\sim x^2\left(m_1\right),\frac{Q_2}{\sigma^2}\sim x^2\left(m_2\right),\cdots,\frac{Q_k}{\sigma^2}\sim x^2\left(m_k\right)$$

Cochranの定理より

- 二つの確率変数
  - ただしm>k

$$X_1^{\,2} \sim x^2(m), X_2^{\,2} \sim x^2(k)$$

  - 差が非負定値

$$X^2 = X_1^{\,2} - X_2^{\,2} \geq 0$$

  - なりたつ

$$X^2 \sim x^2(m-k)$$