応用システム工学
第三回　確率統計の基礎

平成24年05月21日
多変数の確率分布
一般化線形モデル
相関分析
回帰分析

用語

- 測定値または観測値（原因）
  - 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:結果)
  - 目的変数
  - 反応変数
  - 結果変数
  - 従属変数

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

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

  - $X_1, X_2, \cdots, X_n$ の同時確率密度関数

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

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(m<n)

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

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

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

確率変数の独立性1

- 二つの確率変数X,Y
- 独立性の定義

$$P(X \in A \text{かつ} Y \in B) = P(X \in A)P(Y \in B)$$
$$\Leftrightarrow f_{XY}(x,y) = f_X(x) f_Y(y)$$

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

$$P(X_1 \in A_1 \text{かつ} X_2 \in A_2 \text{かつ} \cdots \text{かつ} X_n \in A_n)$$
$$= P(X_1 \in A_1)P(X_2 \in A_2)\cdots P(X_n \in A_n)$$
$$\Leftrightarrow f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

- 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

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

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

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　$Z = z$

    $\Leftrightarrow X + Y = z$

    $\Leftrightarrow Y = z - x$かつ$X = x$
  - 離散分布　$X \in M$
  - Z=zとなる確率

X,Yが独立の時，Y=z-xかつX=xとなる確率

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

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

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

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

線形関数

  - 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

多変量解析　⇒　相関分析，回帰分析

相関解析
相関係数

- 二つの確率変数の間の類似の度合い（相関関係の強さ）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ρxy

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

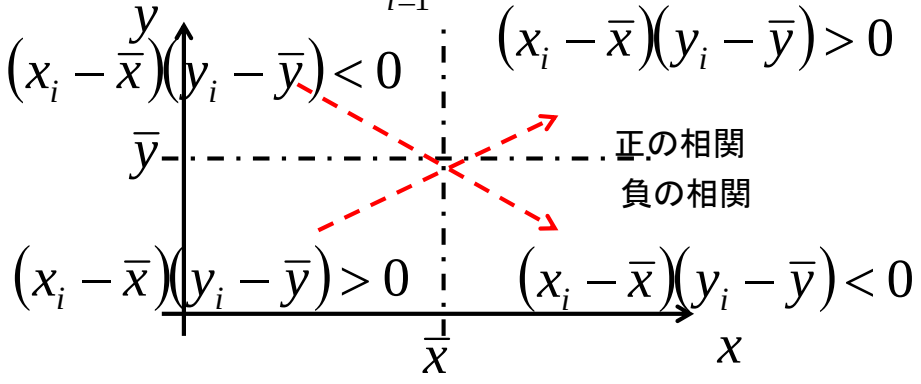
$\bar{x}, \bar{y}$ は各々の相加平均

$$|\rho_{xy}| < 1$$

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

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



回帰分析とは

- データ
  - 目的変数y
  - 目的変数に影響をおよぼす説明変数x
- 分析
  - 予測式を求める
    （ある変量の変化をもう一方の変量の変化で説明するための関数を求める）

$$a_0 + a_1 x_1 + \cdots + a_p x_p \rightarrow y$$

線形回帰（直線回帰）

- 説明変数の数による回帰分析の分類
  - 1個 → 単回帰分析
  - 2個以上 → 重回帰分析

n個のデータの
目的変数yと
説明変数xの組

例
地区数n
地区iの
世帯数x_i, ごみの量y_i

	目的変数:y	説明変数:x
1	y_1	x_1
2	y_2	x_2
…		
i	y_i	x_i
…		
n	y_n	x_n

線形回帰モデル

- モデル式

$$y_i = a_0 + a_1 x_i + e_i \quad (i = 1,2,\cdots,n)$$

  - 未知の定数 a_0, a_1
  - 予測誤差 e_i
- 予測誤差が最小となる定数 $\hat{a}_0, \hat{a}_1$ を求める
  - 予測誤差の平方和を最小にする

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_i)\}^2$$

    - 最小二乗法→ $\hat{a}_0, \hat{a}_1$