第 2 回　単回帰分析に必要な分析道具

[1] シグマ記号による平均と偏差の表現

n 個の数 $x_1 \cdots x_n$ の平均：　$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$　⇒　平均 $\bar{x}$ からの「偏差」：　$x_i - \bar{x}$

[2] 微分

(1) 常微分 $\frac{dy}{dx}$ ：関数 $y = f(x)$について，x の微小な変化に対する y の変化の割合（傾き）

（公式）　① $y = x^c \quad \Rightarrow \quad \frac{dy}{dx} = cx^{c-1}$　　② $\frac{d[f(x)+g(x)]}{dx} = \frac{df(x)}{dx} + \frac{dg(x)}{dx}$

③ $\frac{df(g(x))}{dx} = \frac{df(z)}{dz}\frac{dg(x)}{dx}$　　$y = f(z)$，　$z = g(x)$

(2) 偏微分 $\frac{\partial y}{\partial x_i}$ ：多変数関数 $y = f(x_1, x_2, \cdots, x_n)$ について，x_j $(j \neq i)$ を固定するとき，

x_i の微小な変化に対する y の変化の割合

（例）$y = x_1^2 + x_1x_2 + x_2^2 - 4x_1 - 2x_2$ のとき，　$\frac{\partial y}{\partial x_1} = 2x_1 + x_2 - 4$，　$\frac{\partial y}{\partial x_2} = x_1 + 2x_2 - 2$

(3) 最大化（最小化）問題への応用

(a) 関数 $y = f(x)$の最大化（最小化）問題

最大化（最小化）のための条件：　$\frac{df(x)}{dx} = 0$

(b) 多変数関数　$y = f(x_1, x_2, \cdots, x_n)$ の最大化（最小化）問題

最大化（最小化）のための条件：　すべての i について $\frac{\partial f(x_1, \cdots, x_n)}{\partial x_i} = 0$

（例）$y = x_1^2 + x_1x_2 + x_2^2 - 4x_1 - 2x_2$ の最小値を求めるには，$\frac{\partial y}{\partial x_1} = 0$　と　$\frac{\partial y}{\partial x_2} = 0$

より，$x_1 = 2$，　$x_2 = 0$ であり，y の最小値は -4　となる。

[3] 連続確率変数とは？

・確率変数：どの値が実現するかわからないが，実現する値の範囲がわかっており，
それぞれの値（値の範囲）が実現する可能性を確率で表現できる変数。

1) 離散確率変数：有限個の値をとる確率変数

⇒ 確率変数がとるそれぞれの値について，確率を計算できる。

2) 連続確率変数：無限個の実数値をとる確率変数

⇒ 確率変数がとる範囲について，確率を計算する。

・密度関数 $f(x)$ は連続確率変数 X の現れ方を表し，次の性質をもつ。

① 任意の x について $0 \le f(x) \le 1$ （$f(x)$：$X = x$ のときの確率密度）

② x がとるすべての値について，$\int f(x)dx = 1$ （確率密度の合計は 1 に等しい）

例えば，X の密度関数 $f(x)$ がベル型の曲線であるとき，X が a と b の間に入る確率：$P(a \le X \le b)$ は次の図のように表せる。

・累積分布関数 $F(a)$ は連続確率変数 X の確率を計算するのに役立つ。

$$F(a) = P(X \le a) = \int_{-\infty}^{a} f(x)dx$$

よく使う確率変数の分布（正規分布，t 分布，χ^2 分布，F 分布）については，累積分布関数の値は数表化されており，その表を使って，次のような確率が計算できる。

・ $P(X \ge b) = 1 - F(b)$

・ $P(a \le X \le b) = F(b) - F(a)$

[4] 一つの連続確率変数の特徴

平均：分布の中心の尺度：　$E(X) = \int xf(x)dx$

分散：分布の広がりの尺度：　$Var(X) = E[(X - E(X))^2]$　　※ **標準偏差**：分散の平方根

（平均だけが異なる確率変数 X，Y の分布）　（分散だけが異なる確率変数 X，Y の分布）

（期待値と分散の計算）a と b が定数のとき，確率変数 X の 1 次関数 $Y = aX + b$ について

$E(Y) = E(aX + b) = aE(X) + b$　　$Var(Y) = Var(aX + b) = a^2Var(X)$

[5] 複数の連続確率変数の特徴

(1) 共分散：二つの確率変数の「線形関係の強さ」の尺度

$$Cov(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E(XY) - E(X)E(Y)$$

（解釈例）$Cov(X, Y) > 0 \Rightarrow$ X がふえると Y もふえる。

(2) 相関係数：二つの確率変数の「線形関係の強さ」の尺度を −1 から 1 の値で表したもの。

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \quad (-1 \le Corr(X, Y) \le 1)$$

$Corr(X, Y) > 0$：正の相関，　$Corr(X, Y) < 0$：負の相関，　$Corr(X, Y) = 0$：無相関

(3) 二つの確率変数の同時密度関数と確率的独立性

一般に，確率変数 X と Y は正または負の相関をもち，その現れ方は同時密度関数 $f(x, y)$ で表される。同時密度関数が $f(x, y) = g(x)h(y)$ と書けるとき，X と Y は「確率的に独立」という。X と Y が「確率的に独立」であれば，X と Y は「無相関」である（逆は不成立）。

[6] 統計的推測（推定）の基本

(1) 統計的推測（推定）とは？

・**母集団**：興味のある事象の全体　⇒　特徴は**パラメータ**（例：平均）で表される。

・ **標本** ：母集団から抽出される一部の事象

ランダム標本：ある母集団から「確率的に独立に」抽出された標本

・**統計的推測**：標本の特性から母集団の特性（パラメータ）の値を知ろうとすること。

（例）日本のサラリーマン全員についての平均年収を知りたい場合

母集団：サラリーマン全員の年収　　パラメータ：全員の平均年収

↓　　↑　**統計的推測（推定）**

標　本：サラリーマン千人の年収　→　標本の特性：千人の平均年収

(2) 望ましい統計的推測（推定）の方法とは？

・**推定量**：標本を使って興味のあるパラメータを計算する方法

（例）平均 m の母集団からとったランダム標本を $Y_1,\cdots,Y_n$ とするとき，

パラメータ m の「推定量」の例　$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$

標本 {2, 3, 5, 7, 8} ⇒ m の「推定値」$=\frac{2+3+5+7+8}{5}=5$

・推定量 W でパラメータ θ を推定するとき，推定量の望ましい性質は次のようである。

① **不偏性**：　$E(W)=\theta$

（意味）ランダム標本を何度もとって W を計算すれば，W は平均的に θ に等しい。

② **効率性**：不偏性をもつ推定量の中で，分散が最も小さい（推定精度がよい）。