

# Parametric Probability Density Estimation

Toru Wakahara

## [1] 確率密度関数の推定

- パラメトリックな方法

## [2] 代表例としての正規分布

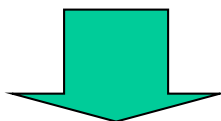
- 正規分布の重要な性質
- いくつかの識別関数の構成法

## [3] 正規分布パラメータの推定

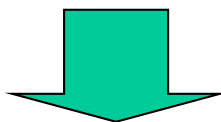
- 最尤法
- ベイズ推定法

# 確率密度関数推定の目的

有限の学習データ:  $\{\mathbf{x}_n, C_n\} (1 \leq n \leq N)$



クラス条件付確率密度関数:  $p(\mathbf{x} | C_k) (1 \leq k \leq c)$



$$P(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k) P(C_k)}{\sum_{j=1}^c p(\mathbf{x} | C_j) P(C_j)} \rightarrow \text{max for } k$$

事後確率 識別

# 確率密度関数推定の方法

## [1] パラメトリックな方法

確率密度関数として特定の関数形を選択し、関数パラメータをデータセットに合うように最適化する。

## [2] ノンパラメトリックな方法

自由度の大きな汎用関数を用いて、データセットを忠実に反映する確率密度関数を構成する。

## [3] セミパラメトリックな方法

特定の関数形の重ね合せ（混合分布）でデータセットの分布を近似する。

# パラメトリックな方法－正規分布

1次元の正規分布:

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

平均:  $\mu = E[x] = \int_{-\infty}^{\infty} xp(x) dx$

分散:  $\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx$

# パラメトリックな方法－正規分布

$d$ 次元の正規分布:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

平均:  $\boldsymbol{\mu} = E[\mathbf{x}]$

共分散行列:  $\Sigma = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t]$

パラメータの数は?

$\Sigma \rightarrow d(d+1)/2, \boldsymbol{\mu} \rightarrow d$       合計  $d(d+3)/2$  個

# 正規分布の等高線 (1)

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

*Mahalanobis distance*       $\Delta^2 = \text{定数である点 } x \text{ では}$   
 $\text{確率密度 } p(x) \text{ が等しい!}$

$\Delta^2 = \text{定数である点 } x \text{ はどんな分布をしているのか?}$   
→ 超楕円体

これを確認するために共分散行列  $\boldsymbol{\Sigma}$  の固有ベクトル,  
固有値を求める

$$\boldsymbol{\Sigma} \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

$$(\mathbf{u}_i, \mathbf{u}_j) = \delta_{ij}$$

$\mathbf{u}_i$  : 固有ベクトル (実数)

$\lambda_i$  : 固有値 (実数)

$$\lambda_1 \geq \dots \geq \lambda_i \geq \dots \geq \lambda_d \geq 0$$

## 正規分布の等高線 (2)

$\Sigma$  の対角化を行う  $\rightarrow$  主軸変換

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i \quad \Leftrightarrow \quad \Sigma \mathbf{U} = \mathbf{U} \mathbf{D}$$

但し

$$\mathbf{U} = (\mathbf{u}_1 \mathbf{u}_2 \dots \mathbf{u}_i \dots \mathbf{u}_d), \quad \mathbf{D} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix}$$

主軸変換行列 固有値  
行列

$$\therefore \mathbf{U}^{-1} \Sigma \mathbf{U} = \mathbf{U}^t \Sigma \mathbf{U} = \mathbf{D}, \quad \Sigma^{-1} = \mathbf{U} \mathbf{D}^{-1} \mathbf{U}^t$$

# 正規分布の等高線 (3)

$\Sigma$  の固有ベクトル, 固有値を用いて主軸変換を行う

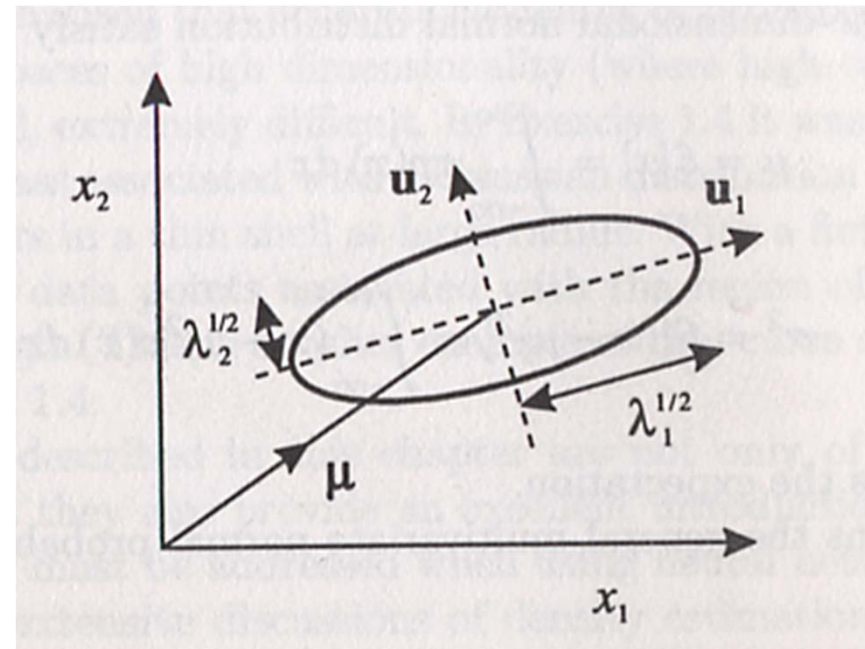
$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^d \frac{(x_i^* - \mu_i^*)^2}{\lambda_i} \quad \text{: 超楕円体}$$

但し

$$\mathbf{x}^* = U^t \mathbf{x}, \quad \boldsymbol{\mu}^* = U^t \boldsymbol{\mu}$$

$$U = (\mathbf{u}_1 \mathbf{u}_2 \dots \mathbf{u}_i \dots \mathbf{u}_d)$$

主軸変換行列



# 正規分布の重要な性質

1. 解析的に扱い易い

2. 中心極限定理:

独立な  $M$  個の確率変数の和の分布は

$M \rightarrow \infty$  で正規分布に近づく!

3. *Mahalanobis distance* は座標軸の正則変換で

2次形式を保存  $\rightarrow$  正規分布のまま

4. 周辺 (*marginal*) 分布も正規分布になる

5.  $\Sigma$  は対角化可能である

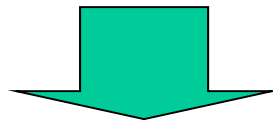
$\rightarrow$  変数が独立に

$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i)$$

# 識別関数の構成

$$p(\mathbf{x} | C_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^t \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}$$

→ 識別関数  $y_k(\mathbf{x}) = \ln p(\mathbf{x} | C_k) + \ln P(C_k)$  に代入



$$y_k(\mathbf{x}) = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^t \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \ln |\Sigma_k| + \ln P(C_k)$$

$$y_k(\mathbf{x}) > y_j(\mathbf{x}) \quad \text{for } \forall j \neq k$$

→  $\mathbf{x}$  is assigned to class  $C_k$

# 識別関数の簡単化(その1)

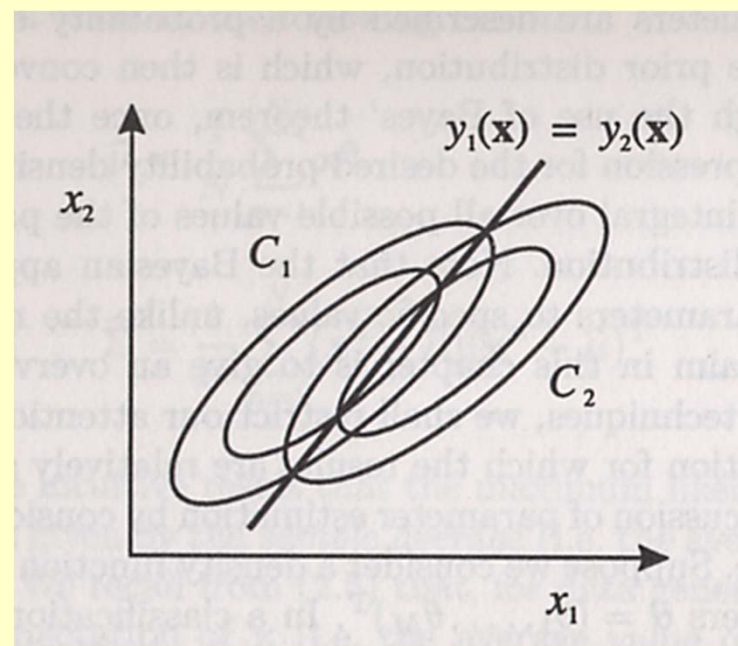
$$y_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| + \ln P(C_k)$$

1)  $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$  の場合

$$y_k(\mathbf{x}) = \mathbf{w}_k^t \mathbf{x} + w_{k0}$$

$$\mathbf{w}_k^t = \boldsymbol{\mu}_k^t \boldsymbol{\Sigma}^{-1}$$

$$w_{k0} = -\frac{1}{2} \boldsymbol{\mu}_k^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln P(C_k)$$

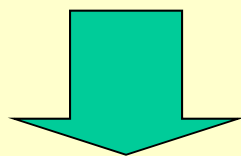


決定境界は超平面

# 識別関数の簡単化(その1)

$\Sigma_k = \Sigma$  の場合の決定境界

$$y_1(\mathbf{x}) = \mathbf{w}_1^t \mathbf{x} + w_{10} = \mathbf{w}_2^t \mathbf{x} + w_{20} = y_2(\mathbf{x})$$



$$\mathbf{w}^t (\mathbf{x} - \mathbf{x}_0) = 0$$

← 決定境界は超平面

$$\mathbf{w} = \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$\mathbf{x}_0 = \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{\ln \frac{P(C_1)}{P(C_2)}}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

## 識別関数の簡単化(その2)

$$y_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| + \ln P(C_k)$$

2)  $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$  の場合

$$y_k(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_k\|^2}{2\sigma^2} + \ln P(C_k)$$

さらに  $P(C_k) = 1/c$  ならば

→ *prototypes*  $\{\boldsymbol{\mu}_k\}$  による最近傍決定則

# パラメトリックな確率密度関数の推定

確率密度関数の関数形の選択

→ パラメータの最適化

## (1) 最尤法

学習データを用いて、尤度を最大化するパラメータを1つ決定する.

## (2) ベイズ推定法

学習データとベイズの定理を用いて、パラメータの確率分布を求める.

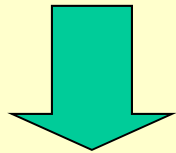
# 最尤法によるパラメータ推定

確率密度関数:  $p(\mathbf{x} | \boldsymbol{\theta})$ ,  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_M)^t$

学習データ:  $X = \{ \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \}$

$\boldsymbol{\theta}$ の尤度:

$$p(X | \boldsymbol{\theta}) = \prod_{n=1}^N p(\mathbf{x}_n | \boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}) \rightarrow \text{max for } \boldsymbol{\theta}$$

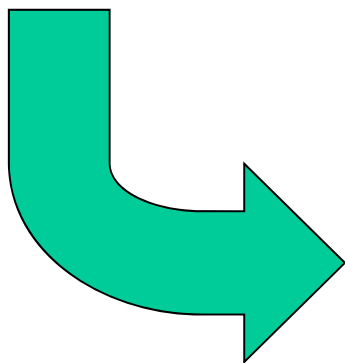


$$E = -\ln L(\boldsymbol{\theta}) = -\sum_{n=1}^N \ln p(\mathbf{x}_n | \boldsymbol{\theta}) \rightarrow \text{min for } \boldsymbol{\theta}$$

# 正規分布の場合の最尤法

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$$

$$E = -\ln L(\boldsymbol{\mu}, \Sigma) = -\sum_{n=1}^N \ln p(\mathbf{x}_n | \boldsymbol{\mu}, \Sigma) \rightarrow \text{min for } \boldsymbol{\mu}, \Sigma$$



$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \hat{\boldsymbol{\mu}})(\mathbf{x}_n - \hat{\boldsymbol{\mu}})^t$$

# 1次元の正規分布の場合

1次元の正規分布:

$$p(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

に従う  $N$  点のデータのマイナス対数尤度:

$$E = -\ln L(\mu, \sigma^2) = -\sum_{n=1}^N \ln p(x_n | \mu, \sigma^2)$$

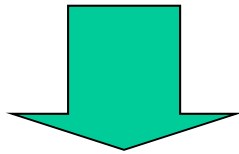
の最小化により次式が得られることを証明せよ.

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2$$

# 最尤法の欠点 (1)

1次元正規分布:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2$$



最尤法で得られる  $\hat{\mu}, \hat{\sigma}^2$  は母集団の  $\mu, \sigma^2$  の  
不偏推定量になっているだろうか？

期待値を調べて見よう

## 最尤法の欠点 (2)

$$E[\hat{\mu}] = \frac{1}{N} \sum_{n=1}^N E[x_n] = \mu \quad \leftarrow \text{不偏推定量}$$

$$E[(\hat{\mu} - \mu)^2] = E\left[\left\{\frac{1}{N} \sum_{n=1}^N (x_n - \mu)\right\}^2\right]$$

$$= \frac{1}{N^2} E\left[\left\{\sum_{n=1}^N (x_n - \mu)\right\}^2\right] = \frac{\sigma^2}{N}$$

↑  
標本平均  $\hat{\mu}$  の分散は  $x$  の分散の  $1/N$

## 最尤法の欠点 (3)

$$\begin{aligned} E[\hat{\sigma}^2] &= E\left[\frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2\right] \\ &= E\left[\frac{1}{N} \sum_{n=1}^N \{(x_n - \mu) - (\hat{\mu} - \mu)\}^2\right] \\ &= \frac{1}{N} \sum_{n=1}^N E[(x_n - \mu)^2] - E[(\hat{\mu} - \mu)^2] \\ &= \sigma^2 - \frac{\sigma^2}{N} = \frac{N-1}{N} \sigma^2 \quad \leftarrow \text{不偏でない} \end{aligned}$$

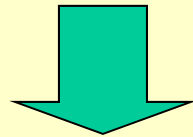
# ベイズ推定法

尤度が最大となる単一の $\theta$ を求めるのではなく、  
 $\theta$ の確率密度関数を求める

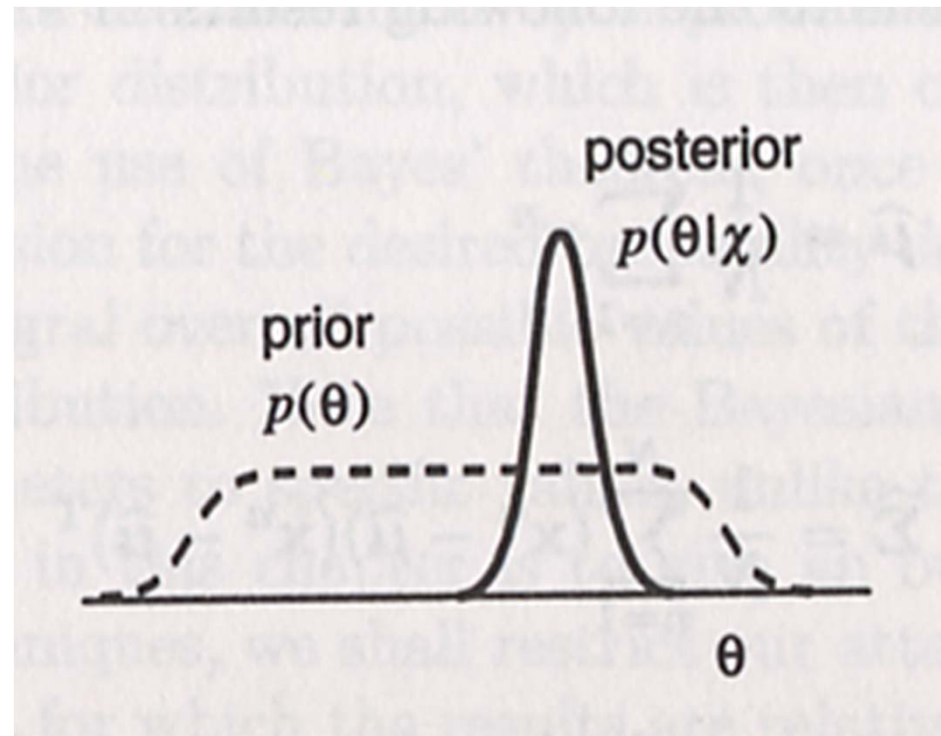
$p(\theta)$ : 事前確率

+

$X$ : データセット



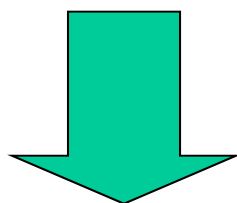
$p(\theta | X)$ : 事後確率



# ベイズ推定法の基本式

ベイズ推定法における  $x$  の確率密度関数:

$$p(\mathbf{x} | \mathbf{X}) = \int p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{X}) d\boldsymbol{\theta}$$



ベイズの定理を用いて  
 $p(\boldsymbol{\theta} | \mathbf{X})$ を推定する

$$p(\boldsymbol{\theta} | \mathbf{X}) = \frac{p(\mathbf{X} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{X})} = \frac{p(\boldsymbol{\theta})}{p(\mathbf{X})} \prod_{n=1}^N p(\mathbf{x}_n | \boldsymbol{\theta})$$

$$p(\mathbf{X}) = \int p(\boldsymbol{\theta}') \prod_{n=1}^N p(\mathbf{x}_n | \boldsymbol{\theta}') d\boldsymbol{\theta}'$$

# ベイズ推定法の適用 (1)

$x$  は次の正規分布 ( $\sigma$  は既知) に従うとする.

$$p(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

一方,  $\mu$  も正規分布に従うと仮定して,  
次式で事前分布  $p_0(\mu)$  ( $\mu_0, \sigma_0$  は既知) を与える.

$$p_0(\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

## ベイズ推定法の適用 (2)

データセット:  $X = \{x_1, x_2, \dots, x_N\}$  が与えられたら,  
 $\mu$  の事後確率  $p_N(\mu | X)$  はどうなるだろうか. ベイズ推定法を適用してみよう.

事後確率: 
$$p_N(\mu | X) = \frac{p_0(\mu)}{p(X)} \prod_{n=1}^N p(x_n | \mu)$$

《ヒント》  $p_N(\mu | X)$  も正規分布になる.  
その平均  $\mu_N$  と分散  $\sigma_N^2$  を求めてみる.

# 1次元の正規分布の場合

未知の  $\mu$  の事前分布を次式で仮定して,

$$p_0(\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp\left\{-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right\}$$

正規分布に従う  $N$  点のデータを用いたベイズ推定より,  
 $\mu$  の事後確率の  $\mu_N, \sigma_N$  が次式となることを示せ.

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \quad \bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$$

## 課題2

具体例として,  $\mu_0 = 0.0, \sigma_0 = \sigma = 0.3$  とする.  
さらに  $\mu$  の真値を  $\mu = 0.8$  として  $\{x_1, x_2, \dots, x_N\}$  を  
正規乱数を用いて発生してみる.  $\mu_N, \sigma_N$  を計算  
して, 下図のようになることを確認しなさい.

### 《ヒント》

#### 正規乱数の発生法

- (1) *Box-Muller*法
- (2) 一様乱数12個の和

