判别分析

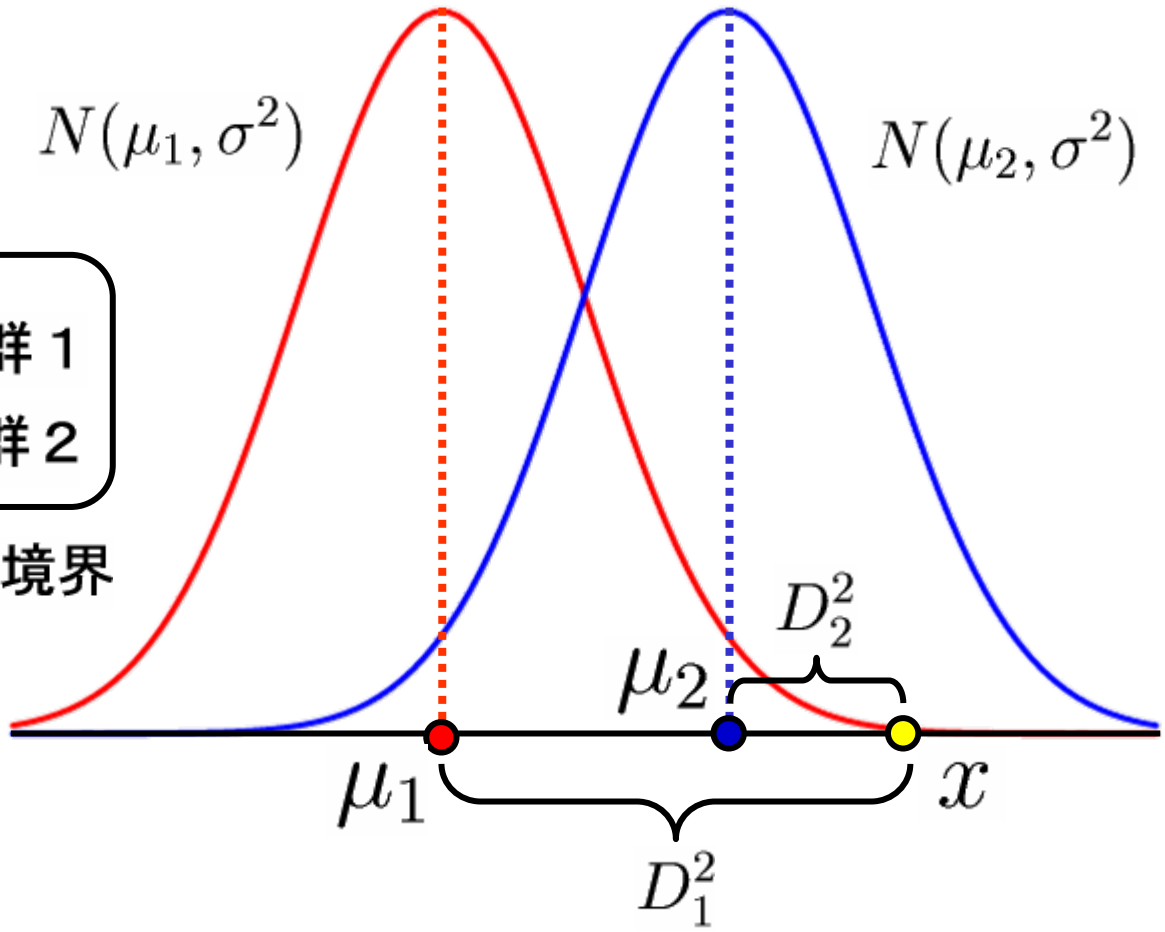
判別問題の例 杉山 髙一著 「多変量データ解析入門」

- A氏、B氏、C氏、D氏の４人のなかの誰かが書いたことは確かである小切手がある。そこに書かれている筆跡から、誰が書いたかを判別する。
- 有価証券報告書にある経常利益率、金利負担率、流動比率、当座比率などにより、企業の倒産を予測する
- 外国の古戦場で発掘した頭蓋骨が、A人種のものか、B人種のものかを判別する。
- 男物着尺地と紳士服地の物性を調べ、布地の特徴をどの物性地が、どの程度よく実現しているかを調べる。

１変量における判別

１変量における判別

左：群1の密度関数
右：群2の密度関数



判別ルール

$D_1^2 < D_2^2 \Rightarrow x \in$ 群１

$D_1^2 > D_2^2 \Rightarrow x \in$ 群２

$D_1^2 = D_2^2$: 判別の境界

誤判別

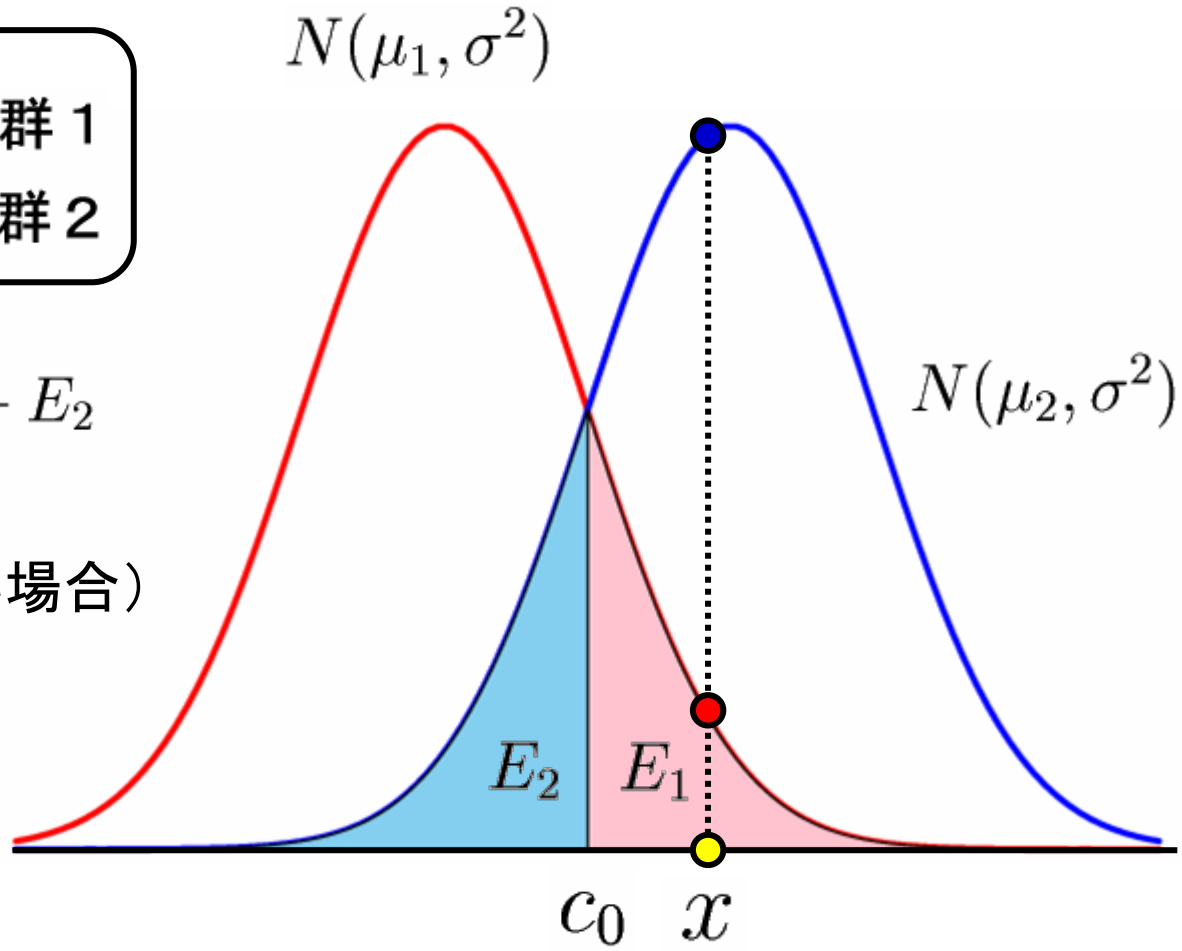
判別ルール

$D_1^2 < D_2^2 \Rightarrow x \in$ 群 1

$D_1^2 > D_2^2 \Rightarrow x \in$ 群 2

- 誤判別率 ： $E_1 + E_2$
- 誤判別率最小
  （母比率が等しい場合）

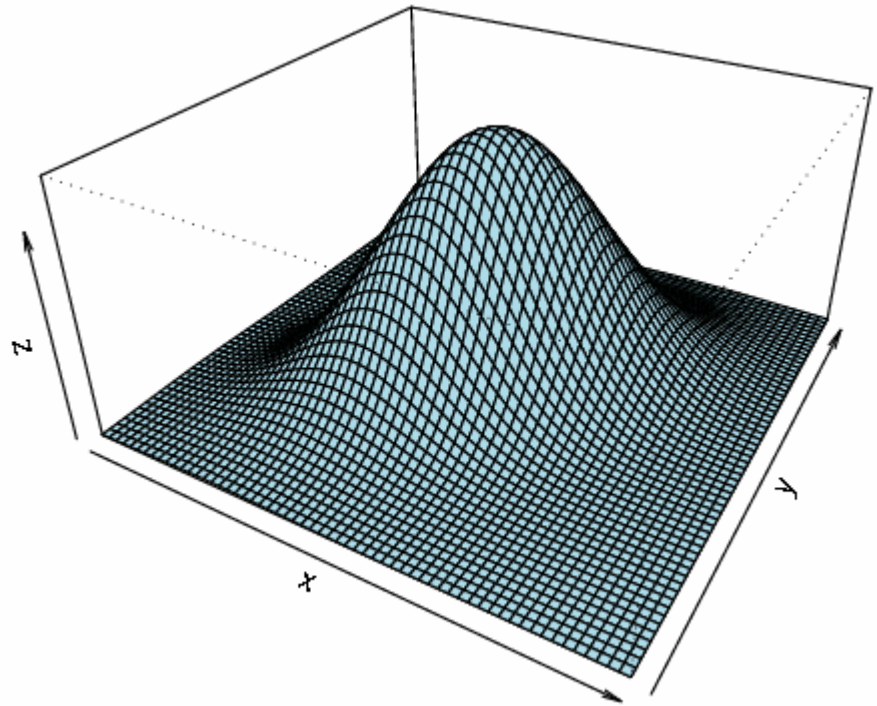
$$c_0 = \frac{\mu_1 + \mu_2}{2}$$



２変量における判別

2変量正規分布

- 2変量正規分布



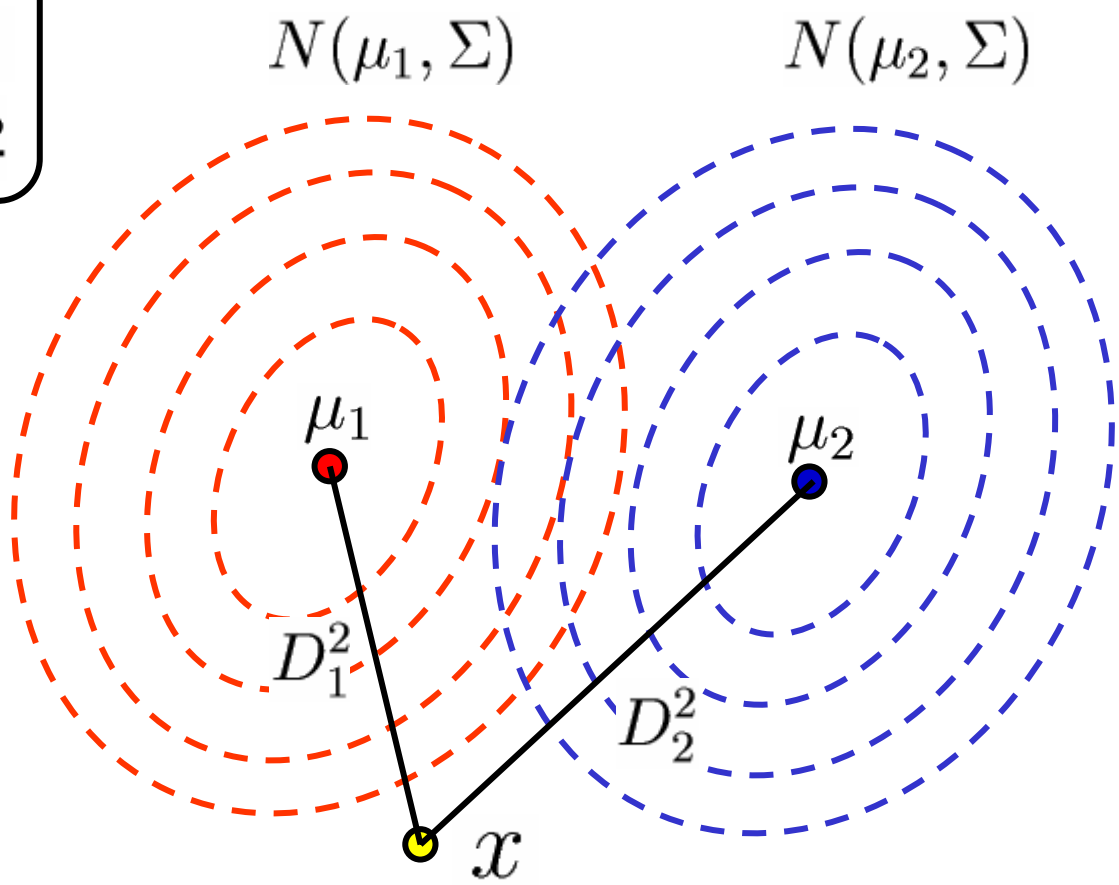
２変量における判別

判別ルール

$D_1^2 < D_2^2 \Rightarrow x \in$ 群１

$D_1^2 > D_2^2 \Rightarrow x \in$ 群２

多変量への拡張も同様



マハラノビス距離

- 今回扱う判別法では、次の3つが必要
  1. 各群の期待値と分散共分散行列
  2. 判別対象の位置
  3. 距離の定義

mean(data)
var(data)

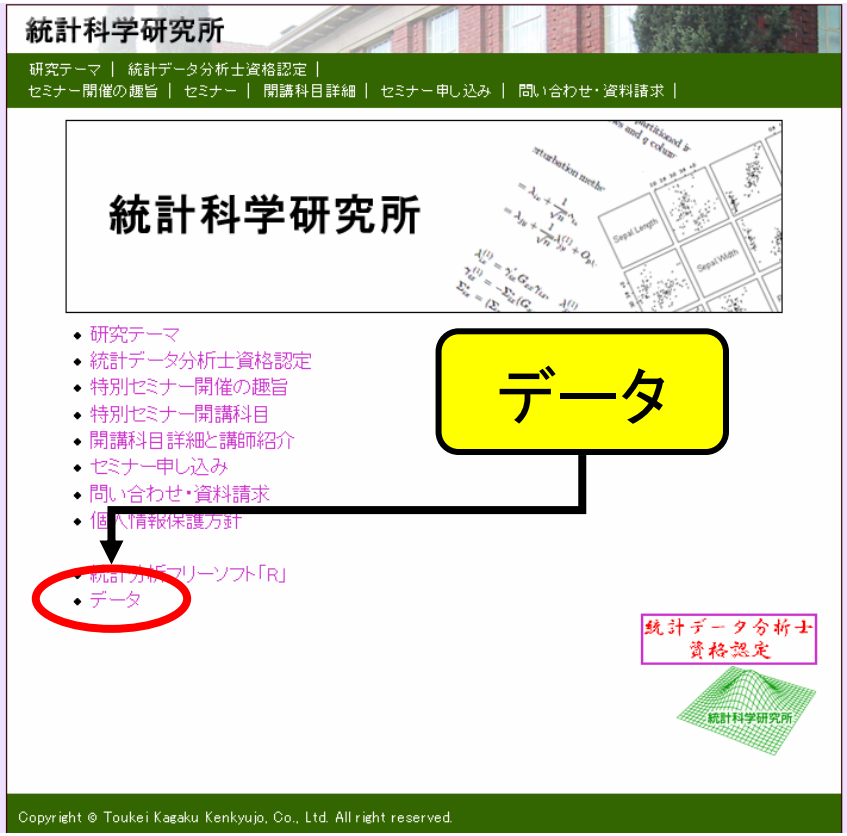
- マハラノビスの距離
  - 確率 ⇔ 距離
  - $D^2 = (\boldsymbol{x} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$
  - $\Sigma = I$ のとき、通常の平方距離

mahalanobis(x, μ, Σ)

筆跡データの分析

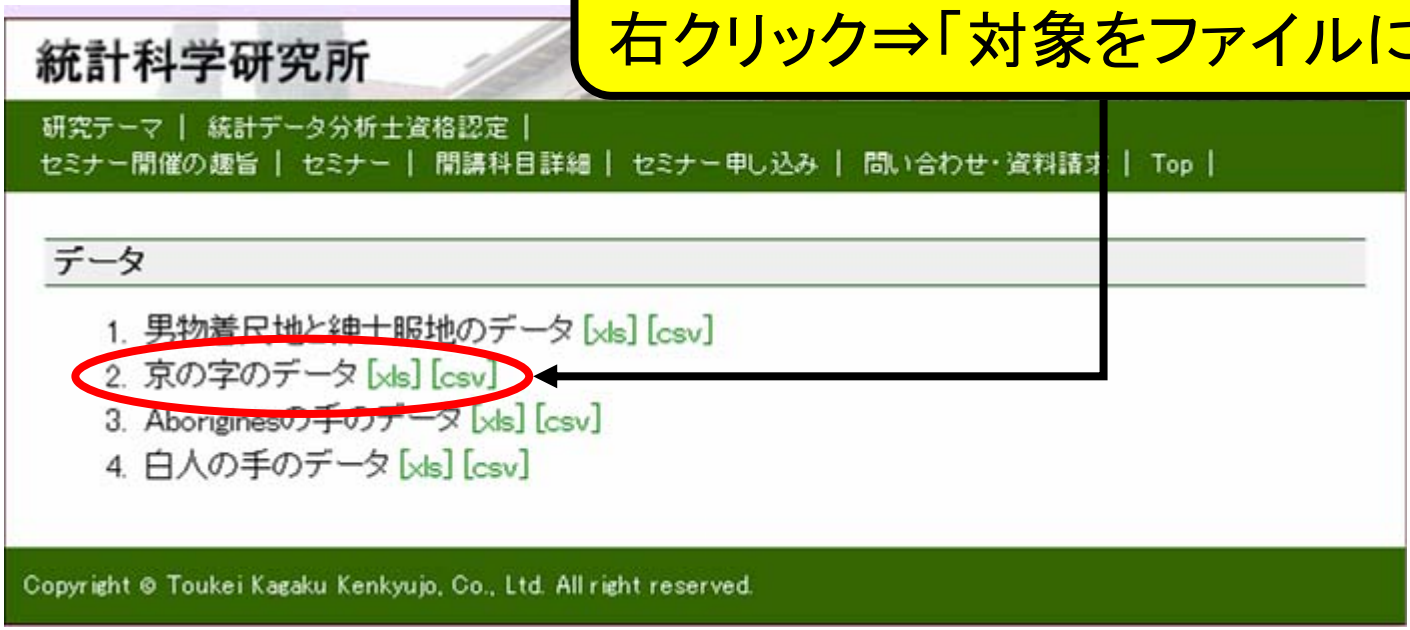
データのダウンロード ①

- 統計科学研究所のウェブサイト
  - http://www.statistics.co.jp/index.htm



データのダウンロード ②

「京の字のデータ」の「csv」を
右クリック⇒「対象をファイルに保存」

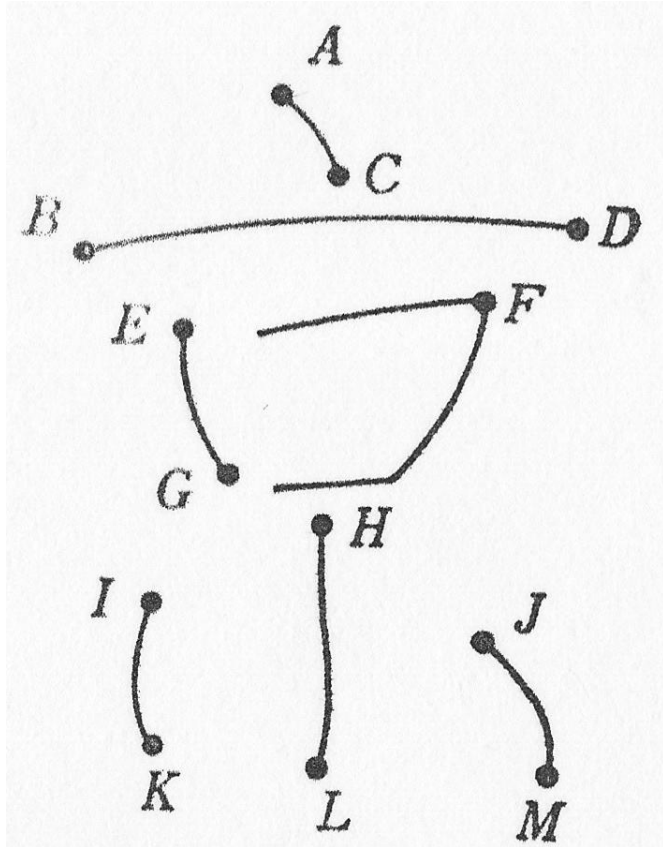


筆跡データの解析

- AかBが書いた「京」の文字がある。
  漢字の筆跡から筆者を識別する問題を考える。
- Aが書いた「京」の字 40字 ： 群1
- Bが書いた「京」の字 40字 ： 群2
- データ ：「京」の字の7箇所を測定
- 目的 ： 群1と群2のデータを分析し、
  新たな「京」の字の筆者を判別する。

筆跡データと変数定義

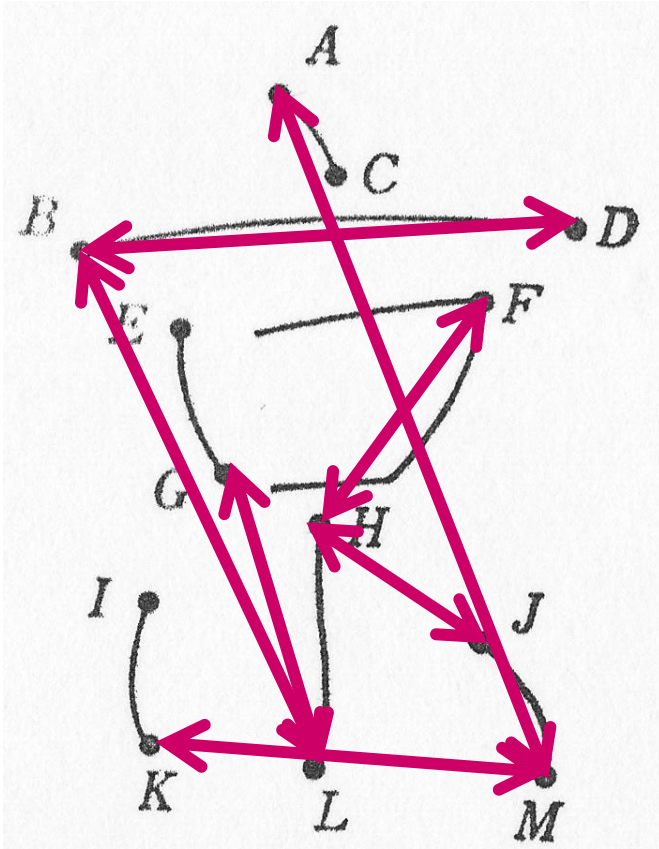
- 変数の定義
  - x_1 : AM/S
  - x_2 : BD/S
  - x_3 : BL/S
  - x_4 : FH/S
  - x_5 : GL/S
  - x_6 : HJ/S
  - x_7 : KM/S



- $\mathrm{S} = \mathrm{AC} + \mathrm{BD} + \mathrm{EG} + \mathrm{EF} + \mathrm{FG} + \mathrm{IK} + \mathrm{HL} + \mathrm{JM}$

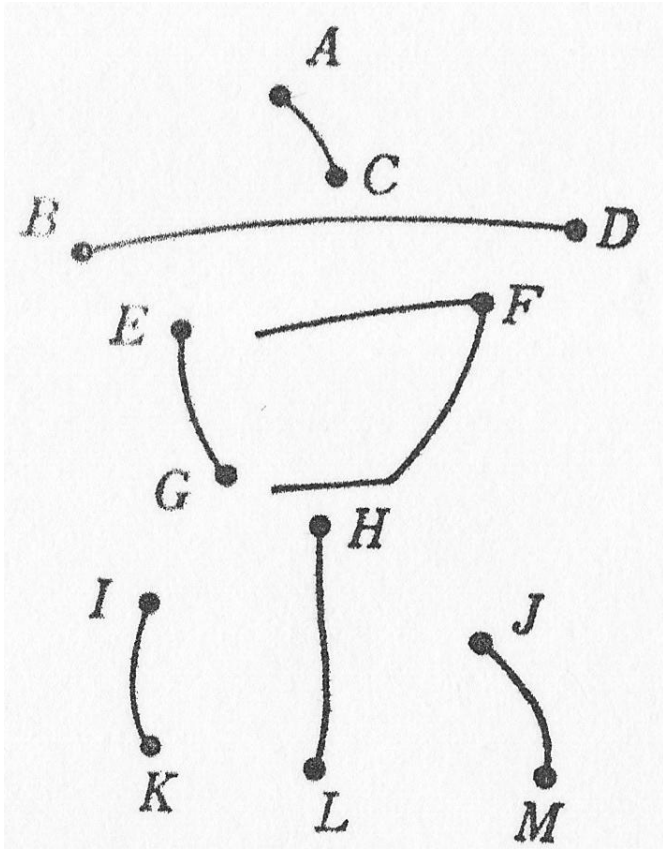
筆跡データと変数定義

- 変数の定義
  - x_1 : AM/S
  - x_2 : BD/S
  - x_3 : BL/S
  - x_4 : FH/S
  - x_5 : GL/S
  - x_6 : HJ/S
  - x_7 : KM/S

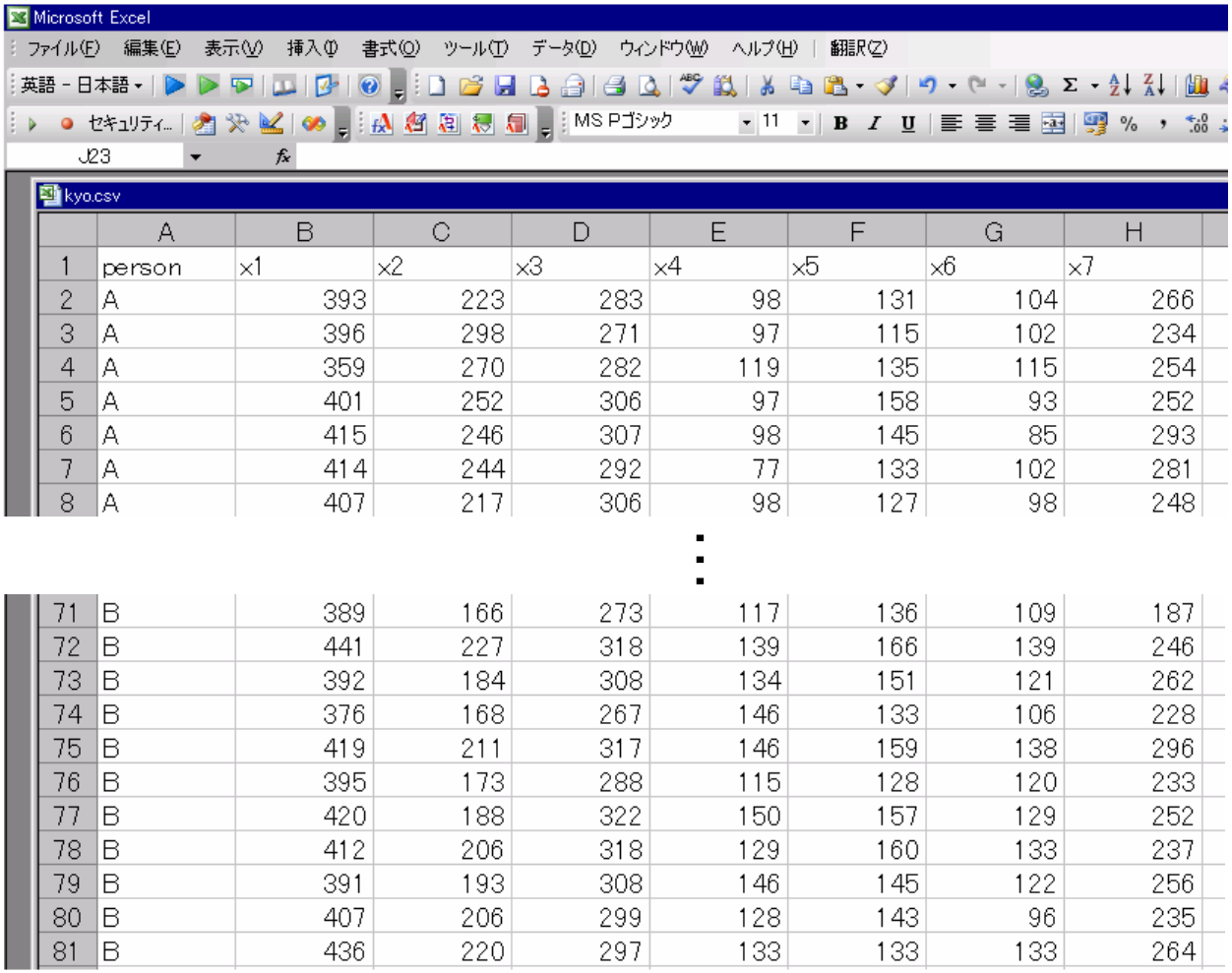


筆跡データ

- ファイル名 ：「kyo.csv」
- 変数名
  - age：年齢
  - x1,...x28：各歯の咬耗度
  - 標本数
    - A：40
    - B：40



「京」の字：筆跡データ



	A	B	C	D	E	F	G	H
1	person	x1	x2	x3	x4	x5	x6	x7
2	A	393	223	283	98	131	104	266
3	A	396	298	271	97	115	102	234
4	A	359	270	282	119	135	115	254
5	A	401	252	306	97	158	93	252
6	A	415	246	307	98	145	85	293
7	A	414	244	292	77	133	102	281
8	A	407	217	306	98	127	98	248
⋮								
71	B	389	166	273	117	136	109	187
72	B	441	227	318	139	166	139	246
73	B	392	184	308	134	151	121	262
74	B	376	168	267	146	133	106	228
75	B	419	211	317	146	159	138	296
76	B	395	173	288	115	128	120	233
77	B	420	188	322	150	157	129	252
78	B	412	206	318	129	160	133	237
79	B	391	193	308	146	145	122	256
80	B	407	206	299	128	143	96	235
81	B	436	220	297	133	133	133	264

プログラム

```
data <- read.csv(“kyo.csv",header=T)
A <- subset(data, person=="A")[,-1]
B <- subset(data, person=="B")[,-1]
n1 <- nrow(A)
n2 <- nrow(B)
m1 <- apply(A, 2, mean)
m2 <- apply(B, 2, mean)
S1 <- var(A)*(n1-1)
S2 <- var(B)*(n2-1)
V <- (S1 + S2)/(n1+n2-2)
mahalanobis(A[1,],m1,V)
mahalanobis(A[1,],m2,V)
```

プログラムの説明 ①

```
A <- subset(data, person==”A”)[,-1]
B <- subset(data, person==”B”)[,-1]
n1 <- nrow(A)
n2 <- nrow(B)
```

- subset(データ, 変数名”==“値”)
  データのサブセットを作成するための関数。
  1行目のプログラムではA氏のデータのみを抽出している。
- nrow(“データ”)
  - データの列数を数える。
    今回扱うデータでは、これが標本数にあたる。

プログラムの説明 ②

```
m1 <- apply(A, 2, mean)
m2 <- apply(B, 2, mean)
```

- apply(データ, 列か行の指定, 適用する関数)
  - 初めの引数は、関数を適用するデータを指定する
  - 2番目の引数に「1」を指定すると各行に、
    「2」を指定すると各列に対して関数を適用する。
  - 3番目の引数には、データに適用したい関数を指定する。

プログラムの説明　③

```
S1 <- var(A)*(n1-1)
S2 <- var(B)*(n2-1)
V <- (S1 + S2)/(n1+n2-2)
```

- S1 <- var(A)*(n1-1) ; S2 <- var(B)*(n2-1)
  - 1行目と2行目では、各群の偏差平方和を求めている。
    ここで、関数「var」は不偏分散として定義されていることに
    注意する。
- V <- (S1 + S2)/(n1+n2-2)
  - 3行目では、共通の分散として、合併共分散行列を求めている。

プログラムの説明 ④

mahalanobis(A[1,],m1,V)
mahalanobis(A[1,],m2,V)

- mahalanobis(A[1,],m1,V)
  - Aの1番目の標本と、1群との
    マハラノビス距離を求めている。
    右の図の D_1^2 を求めている。
- mahalanobis(A[1,],m2,V)
  - Aの1番目の標本と、2群との
    マハラノビス距離を求めている。
    右の図の D_2^2 を求めている。