データの分析と知識発見

Introduction to Data Analysis

今回の構成

Rのデータの形式について学ぶ

ファイルの読み込みについて学ぶ

相関、共分散について学ぶ

Rにおけるデータの型と形式

Rにおけるデータの型

- 数値　0, 1, 2, ...
- 文字列　a, b, c, ...
- 論理値　TRUE, FALSE
- ファクター（第5回）

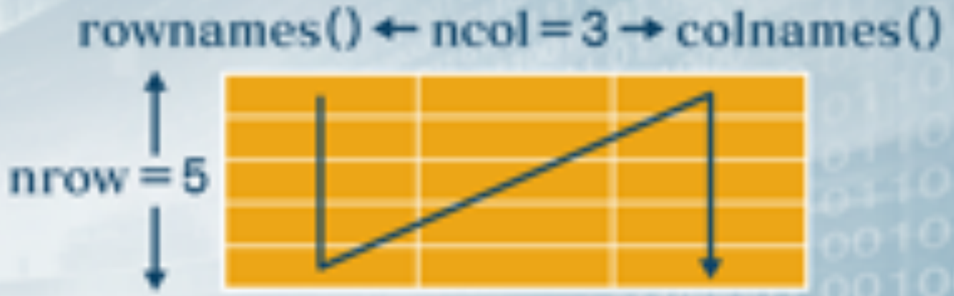
データの形式

- スカラ
- ベクトル
- 行列
- リスト
- データ・フレーム

行列

```
> A <- matrix(c(1,2,3,4,5,6),ncol=2,nrow=3)
> A
     [,1][,2][,3]
[1,]  1  3  5
[2,]  2  4  6
> A[1,2]
[1]3
> A <- matrix(c(1,2,3,4,5,6),ncol=3,byrow=T)
> A
     [,1][,2][,3]
[1,]  1  2  3
[2,]  4  5  6
```

A <- matrix(ベクトル　ncol＝列の数)



- ベクトルの要素はすべて同じもの
- A[1,](1行目)　A[,2](2列目)
- colnames()や rownames()で名前をつけることができる

リスト

```
> sasaki <- list ("英語", "35",c(70,60,40,30))
> names (sasaki) <- c ("科目","年齢","点数")
> sasaki
$ 科目
[1] "英語"
$ 年齢
[1] "35"
$ 点数
[1] 70 60 40 30
```

a <- list (要素1，要素2，要素3)

- 異なる要素をまとめたもの
- a[[2]]というように指定
- 名前をつけた後は$で指定

データフレーム

```
> height <- c (148,152,154,158,163)
> weight <- c (52, 54, 56, 58, 60)
> D <- data.frame (height, weight)
> D
height weight
1  148  52
2  152  54
3  154  56
4  158  58
5  163  60
```

D <- data.frame (要素1, 要素2, 要素3)

- 同じ長さの異なる型のベクトル
- D[1:2,2]というように指定
- D$height などと

ファイルの読み込み

半角スペースやタブ区切り
→ read.table()
カンマ区切り(CSV)
→ read.csv()

1行目が列の名前
→ header＝TRUE
(h＝Tと省略できる)
1列目が各行の名前
→ row.names＝1

ch031.dat
name1A B
A1 148 52
A2 152 54
A3 154 56
A4 158 58
A5 163 60

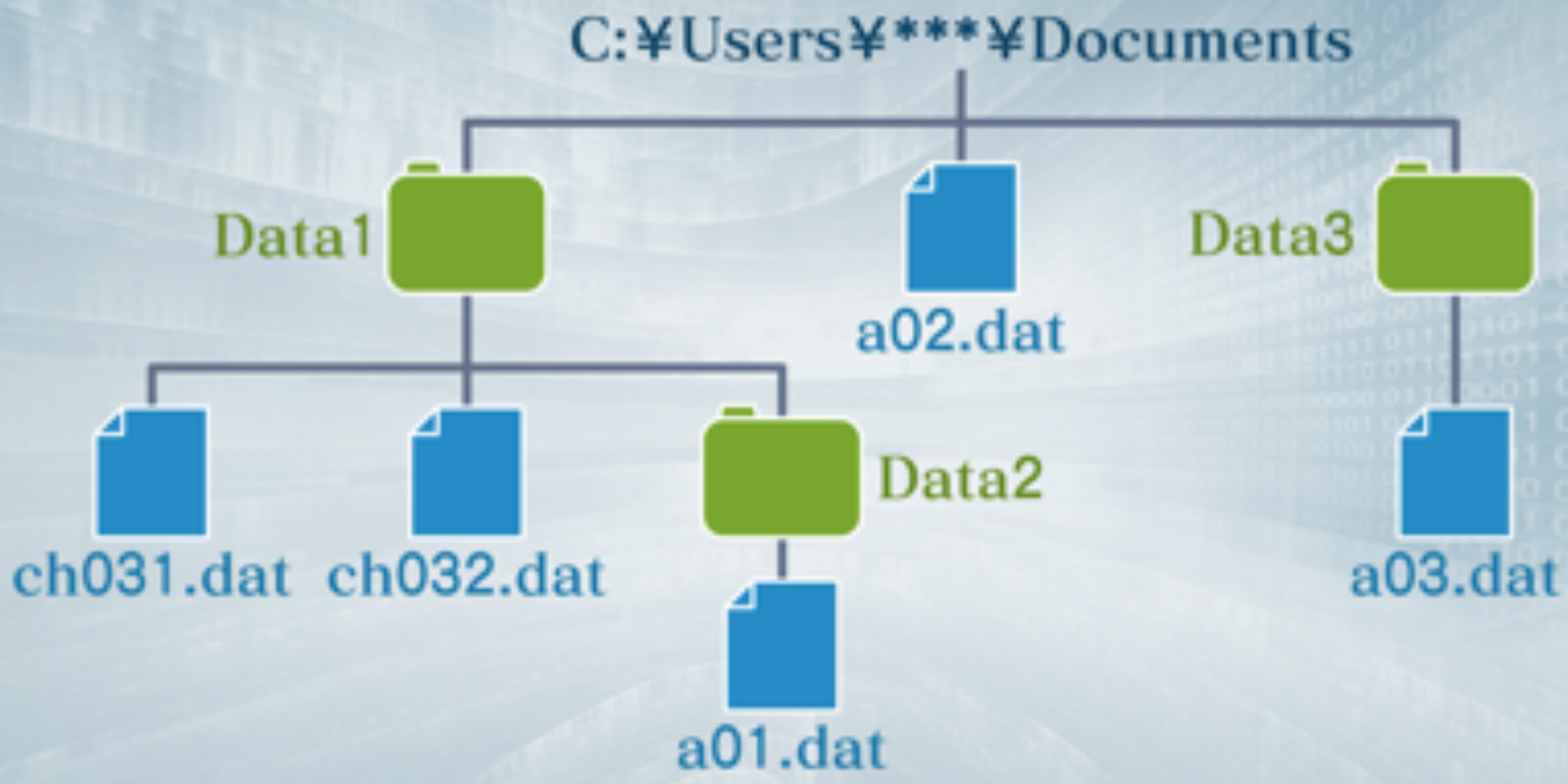
ch032.dat
name1, A, B
A1, 148, 52
A2, 152, 54
A3, 154, 56
A4, 158, 58
A5, 163, 60

Rによる計算

```
> h1 <- read.table("ch031.dat",header=T, row.names=1)
> h2 <- read.table("ch032.dat",header=T, row.names=1, sep=",")
> h3 <- read.csv("ch032.dat",header=T, row.names=1)
```

```
> mean(h1$A)
> var(h1)
> cor(h1)
```

ツリー構造


C:¥Users¥***¥Documents
Data1
a02.dat
Data3
ch031.dat
ch032.dat
Data2
a03.dat
a01.dat

共分散と相関係数

name	x_h	x_w	$x_h-\mu_h$	$x_w-\mu_w$	$(x_h-\mu_h)(x_w-\mu_w)$
A1	148	52			
A2	152	54			
A3	154	56			
A4	158	58			
A5	163	60			
sum	775	280			
ave	155	56			

$\sigma_h^2=33$ $\sigma_w^2=10$

平均と分散

5人の身長と体重

平均

不偏分散

name	x_h	x_w
A1	148	52
A2	152	54
A3	154	56
A4	158	58
A5	163	60
ave	155	56
var	33	10

共分散と相関係数

name	x_h	x_w	$x_h-\mu_h$	$x_w-\mu_w$	$(x_h-\mu_h)(x_w-\mu_w)$
A1	148	52	−7	−4	
A2	152	54	−3	−2	
A3	154	56	−1	0	
A4	158	58	3	2	
A5	163	60	8	4	
sum	775	280	0	0	
ave	155	56	0	0	

偏差を計算

$\sigma_h^2=33$ $\sigma_w^2=10$

共分散と相関係数

name	x_h	x_w	$x_h-\mu_h$	$x_w-\mu_w$	$(x_h-\mu_h)(x_w-\mu_w)$
A1	148	52	−7	−4	28
A2	152	54	−3	−2	6
A3	154	56	−1	0	0
A4	158	58	3	2	6
A5	163	60	8	4	32
sum	775	280	0	0	
ave	155	56	0	0	

共分散を計算

$\sigma_h^2=33$ $\sigma_w^2=10$

共分散と相関係数

name	x_h	x_w	$x_h-\mu_h$	$x_w-\mu_w$	$(x_h-\mu_h)(x_w-\mu_w)$
A1	148	52	−7	−4	28
A2	152	54	−3	−2	6
A3	154	56	−1	0	0
A4	158	58	3	2	6
A5	163	60	8	4	32
sum	775	280	0	0	72
ave	155	56	0	0	18

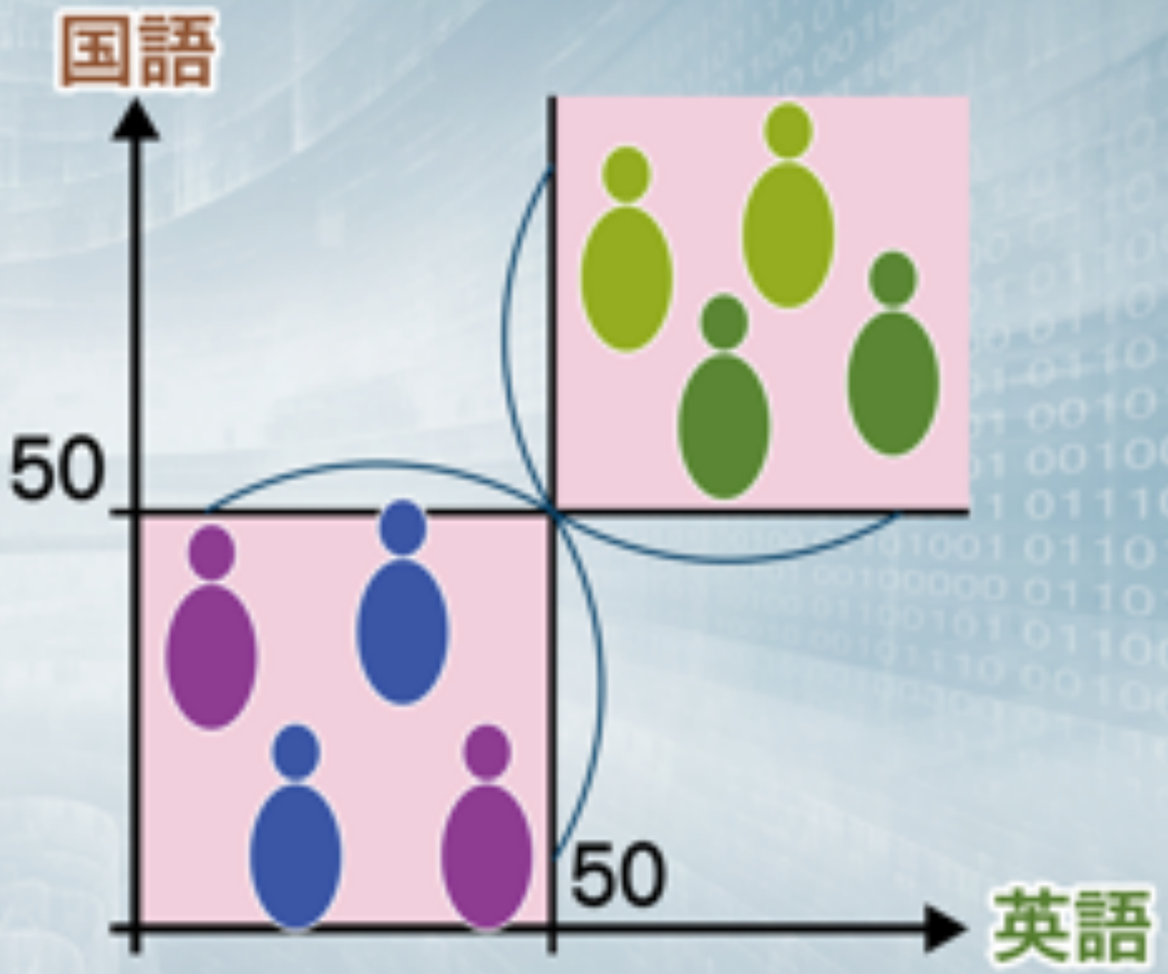
相関係数

$\sigma_h^2=33$ $\sigma_w^2=10$

$$\frac{18}{\sqrt{33}\sqrt{10}}=0.9908674$$

相関係数(1)

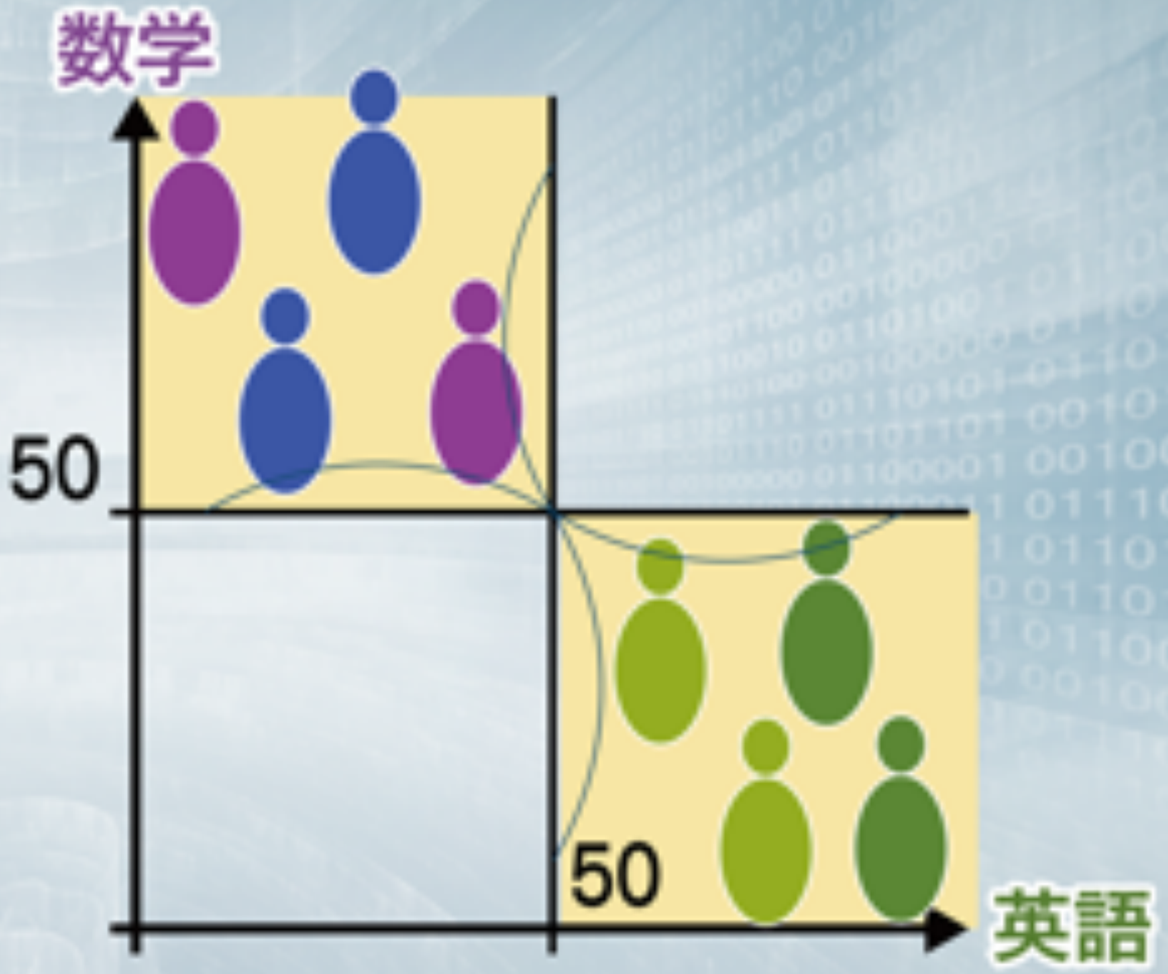
	英語	国語
A	20	15
B	40	15
C	15	40
D	30	45
E	60	25
F	60	60
G	70	85
H	75	65



相関係数(2)

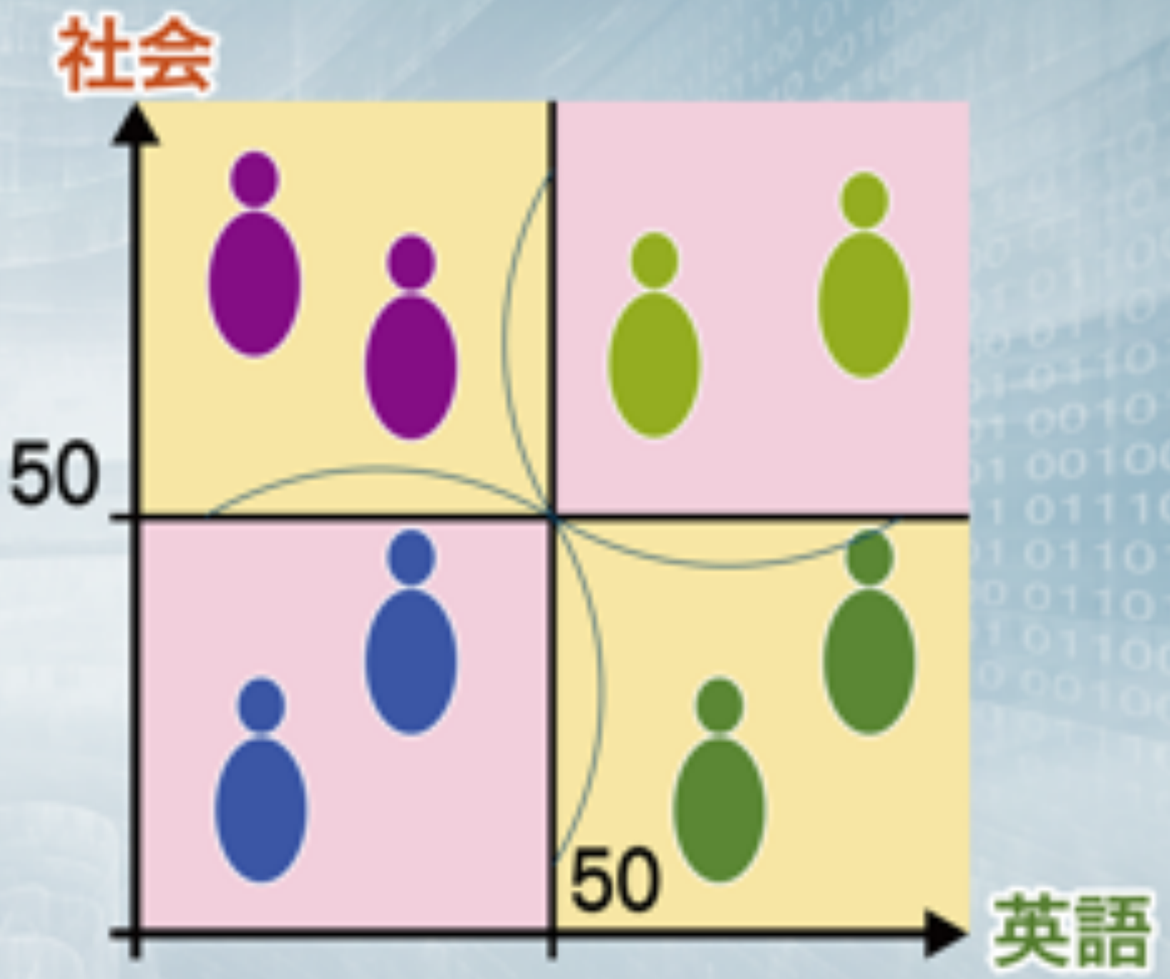
	英語	数学
A	20	55
B	40	60
C	15	85
D	30	90
E	60	40
F	65	10
G	70	45
H	75	10

英語と数学とは
負の関係(相関)がある



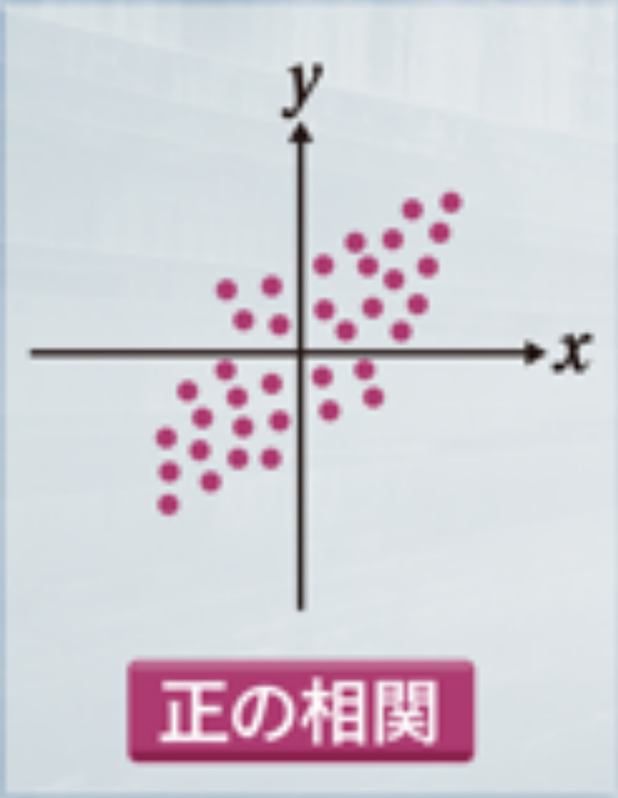
相関係数(3)

	英語	社会
A	20	75
B	40	60
C	20	20
D	40	40
E	60	25
F	70	40
G	55	65
H	70	65



相関係数

相関係数：2つの間の関係の強さを表す



Rによる計算

```
> h1 <- read.table("ch031.dat",header=T, row.names=1)
> h2 <- read.table("ch032.dat",header=T, row.names=1, sep=",")
> h3 <- read.csv("ch032.dat",header=T, row.names=1)
```

```
> mean(h1$A)
> var(h1)
> cor(h1)
```

分散共分散行列と相関行列

分散共分散行列

$$\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1r} \\ s_{21} & s_{22} & \cdots & s_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ s_{r1} & s_{r2} & \cdots & s_{rr} \end{pmatrix}$$

s_{ij} ： i 番目の変数と
j 番目の変数の共分散

$s_{ij} = s_{ij}$

←対称行列→

相関行列

$$\begin{pmatrix} 1 & r_{12} & \cdots & r_{1r} \\ r_{21} & 1 & \cdots & r_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ r_{r1} & r_{r2} & \cdots & 1_{rr} \end{pmatrix}$$

r_{ij} ： i 番目の変数と
j 番目の変数の相関関数

$r_{ij} = r_{ij}$

分散と共分散

N組のデータ $(x_h^{(p)},\ x_w^{(p)})\quad (p=1,2,\cdots,N)$ $\mu_w,\ \mu_w$：平均値

分散 $\sigma_h^2=\dfrac{1}{N-1}\displaystyle\sum_{p=1}^{N}(x_h^{(p)}-\mu_h)^2,\ \ \sigma_w^2=\dfrac{1}{N-1}\displaystyle\sum_{p=1}^{N}(x_w^{(p)}-\mu_w)^2$

共分散

$$s_{hw}=\frac{1}{N-1}\sum_{p=1}^{N}(x_h^{(p)}-\mu_h)(x_w^{(p)}-\mu_w)$$

相関係数

$$r_{hw}=\frac{s_{hw}}{\sigma_h\sigma_w}$$

共分散と相関係数

name	x_h	x_w	$x_h-\mu_h$	$x_w-\mu_w$	$(x_h-\mu_h)(x_w-\mu_w)$
A1	148	52	−7	−4	28
A2	152	54	−3	−2	6
A3	154	56	−1	0	0
A4	158	58	3	2	6
A5	163	60	8	4	32
sum	775	280	0	0	72
ave	155	56	0	0	18

共分散

$\sigma_h^2=33$ $\sigma_w^2=10$

Rによる計算

```
> h1 <- read.table("e01.dat",header=T, row.names=1)
> h2 <- read.table("e02.dat",header=T, row.names=1, sep=",")
> h3 <- read.csv("e02.dat",header=T, row.names=1)
```

```
> mean(h1$A)
> var(h1)
> cor(h1)
```