確率統計と情報処理・演習（2015 年度後期）

2 変量データ分析

2015 年 10 月 09 日

日本女子大学理学部数物科学科　　今野 良彦

August 13, 2015

今日の講義の目的と概要

- ２変量のデータを各変量ごとに「1変量のデータの分析」を適用.
  - ヒストグラム・箱ひげ図・分布の代表値・ばらつき
- 散布図 — 2変量データとして分布を眺めること（変量間の関係）.
- 回帰直線 — 散布図をみて，2変量データが全体として右上がりまたは右下がりの傾向があったときに，その傾向を示す直線を引くこと.
- 相関係数 — ２変量間の関係（直線的な傾向の強弱）を数値で示すもの.
- 順位相関係数 — データが「順位」という特殊な場合の相関係数

身長と体重のデータ

番号	身長	体重	番号	身長	体重
1	148	41	9	137	31
2	160	49	9	149	47
3	159	45	10	160	47
4	153	43	11	151	42
5	151	42	12	157	39
6	140	29	13	157	48
7	156	49	14	144	36

Figure 1: 中学生 14 人の体格データ

cbind

```
> height
 [1] 148 160 159 153 151 140 156 137 149 160 151 157 157 144
> weight
 [1] 41 49 45 43 42 29 49 31 47 47 42 39 48 36
> length(weight)  #オブジェクト weight のデータの個数を調べる.
[1] 14
> length(height)
[1] 14
> bodydata<-cbind(height,weight) #データは行ベクトルを行列配置.
> bodydata
      height weight
 [1,]    148     41
 [2,]    160     49
 [3,]    159     45
 [4,]    153     43
 [5,]    151     42
# 以下は略
> bodydata[1,]
height weight
   148     41
> bodydata[2,]
height weight
   160     49
> bodydata[,1]
 [1] 148 160 159 153 151 140 156 137 149 160 151 157 157 144
```

matrix

```
> bodydata2<-matrix(
+ c(148,160,159,153,151,140,156,137,149,160,151,157,157,144,
+  41,49,45,43,42,29,49,31,47,47,42,39,48,36
+ ),,2)  # matrix(c( ....),列の数,行の数)
> bodydata2
      [,1] [,2]
 [1,]  148   41
 [2,]  160   49
 [3,]  159   45
 [4,]  153   43
 [5,]  151   42
 [6,]  140   29
 [7,]  156   49
 [8,]  137   31
 [9,]  149   47
[10,]  160   47
[11,]  151   42
[12,]  157   39
[13,]  157   48
[14,]  144   36
>
> mean(bodydata[,1])  # 身長の平均
[1] 151.5714
> mean(bodydata[,2])  # 体重の平均
[1] 42
```


グラフの同時表示方法

mfrow

```
> par(mfrow=c(2,1))
> # グラフを 2 行に表示  c(m,n) は m行 n列の意味
> hist(height)
> boxplot(height)
> par(mfrow=c(1,1))  # もとにもどす
```

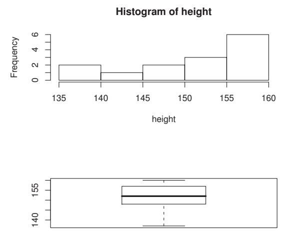

Figure 2: ヒストグラムと箱ひげ図

問題 1

- 身長と体重のデータを行列配置のオブジェクトに入力．ただし，1行目は身長で2行目は体重
- 身長と体重のデータについて，以下ことを求めよ．

(1) ヒストグラム
(2) 平均と中央値
(3) 箱ひげ図（学籍番号-boxplot2.pdf　例：21416***-boxplot2.pdf）
(4) 範囲，四分位(しぶんい)範囲，平均偏差，分散，標準偏差

- 締め切りは 2015 年 10 月 23 日（金）13 時
- 21416***-目白花子-2015-10-23.txt

テキストデータの読み込み

エディターを使い data1.txt を作成する．R の「ファイル」から「ディレクトリの変更」を選択し,「Browse」をクリックして，data1.txt のあるディレクトリを指定する．

data1.txt

```
148 41
160 49
153 45
```

read.table

```
> x<-read.table("data1.txt")
> x
   V1 V2
1 148 41
2 160 49
3 153 45
>
```

データエディタ

data1.txt

```
> bodydata
     height weight
 [1,]    148     41
 [2,]    160     49
 [3,]    159     45
 [4,]    153     43
 [5,]    151     42
 [6,]    140     29
 [7,]    156     49
# 以下は省略
> fix(bodydata)           # データエディタが表示される
> data.entry(bodydata)    # データエディタが表示される
> iris                    # データ iris の表示
> fix(iris)               # データエディタが表示される
>
```


データエディタ

	col1	col2	var3	var4	var5	var6	var7
1	148	41					
2	160	49					
3	159	45					
4	153	43					
5	151	42					
6	140	29					
7	156	49					
8	137	31					
9	149	47					
10	160	47					
11	151	42					
12	157	39					
13	157	48					
14	144	36					
15							
16							
17							
18							
19							

図 4.1　データエディタ

散布図

身長と体重の散布図を作成

data1.txt

```
>
> plot(height,weight)
>
```

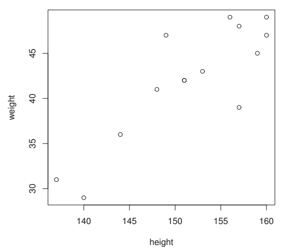

Figure 3: 身長と体重の散布図

散布図から調べることができること

- データは右上がり，右下がり，またはいずれでもない
- 外れ値があるか？

外れ値とは

data1.txt

```
> height2<-c(height,140)
> weight2<-c(weight,60)
> height2
 [1] 148 160 159 153 151 140 156 137 149 160 151 157 157 144 140
> weight2
 [1] 41 49 45 43 42 29 49 31 47 47 42 39 48 36 60
>
> plot(height2,weight2)
```


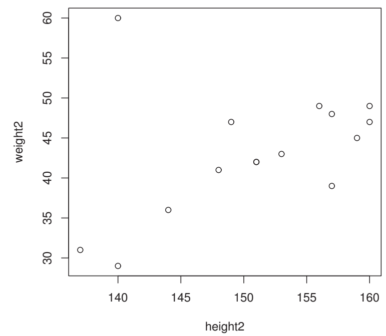

Figure 4: 身長と体重の散布図(外れ値のある場合)

回帰直線

⋆ 散布図を見て，全体が右下がりまたは右上がりの傾向があったとき，その傾向を示す直線

$$y = a + bx$$

を散布図に引くことを考える．

⋆ データに基づいて傾き b と切片 a をどうきめるか？

⋆ 最小自乗法を使う！(他の方法もある！)

最小自乗法

⋆ データを $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ とする．

⋆ $i\,(i = 1, 2, \ldots, n)$ 番目のデータが x_i に対して，

データの y 変量の値	y_i
直線上の y の値	$\hat{y}_i = a + bx_i$
誤差	$y_i - \hat{y}_i$

⋆「誤差」が少ないほどよいと考える．すなわち,「誤差」が零に近いほうどよい!

⋆ 符号の影響をなくして，データ全体の「誤差」を合計する！すわわち

$$Q(a, b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2$$

を最小にする a と b を求める！

最小自乗法

$\star$ a と b について Q を偏微分すると

$$\begin{cases} \frac{\partial Q}{\partial a} = -2\sum_{i=1}^n \{y_i - (a+bx_i)\} = -2n\bar{y}_n + 2an + 2nb\bar{x}_n = 0 \\ \frac{\partial Q}{\partial b} = -2\sum_{i=1}^n x_i\{y_i - (a+bx_i)\} = -2\sum_{i=1}^n x_iy_i + 2na\bar{x}_n + 2b\sum_{i=1}^n x_i^2 = 0 \end{cases}$$

となる．ただし，$\bar{x}_n = (1/n)\sum_{i=1}^n x_i$ と $\bar{y}_n = (1/n)\sum_{i=1}^n y_i$

$$\begin{cases} n\bar{y}_n - an - n\bar{x}_n b &= 0 \\ \sum_{i=1}^n x_iy_i - n\bar{x}_n a - \sum_{i=1}^n x_i^2 b &= 0 \end{cases}$$

よって，

$$\begin{cases} n\bar{x}_n\bar{y}_n - n\bar{x}_n a - n\bar{x}_n^2 b = 0 \\ \sum x_iy_i - n\bar{x}_n a - \sum x_i^2 b = 0 \end{cases} \tag{1}$$

よって，$\sum_{i=1}^n x_i^2 - n\bar{x}_n^2 \neq 0$ ならば，

$$\begin{aligned} b &= \frac{\sum_{i=1}^n x_iy_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^n x_i^2 - n\bar{x}_n^2} = \frac{\sum_{i=1}^n (x_i-\bar{x}_n)(y_i-\bar{y}_n)}{\sum_{i=1}^n (x_i-\bar{x}_n)^2} \\ &= \frac{(1/n)\sum_{i=1}^n (x_i-\bar{x}_n)(y_i-\bar{y}_n)}{(1/n)\sum_{i=1}^n (x_i-\bar{x}_n)^2} \\ &=: \frac{s_{xy}}{s_x^2} \\ a &= \bar{y}_n - b\bar{x}_n = \bar{y}_n - \frac{s_{xy}}{s_x^2}\bar{x}_n \end{aligned} \tag{2}$$

以後は，$\hat{b} = s_{xy}/s_x^2$ と $\hat{a} = \bar{y}_n - \hat{b}x_n$ と書くことにする．(データから求めた回帰直線の傾き $\hat{b}$ と切片 $\hat{a}$ という意味)

データより求めた回帰直線の方程式

$s_x^2 \neq 0$ のとき，

$$y - \bar{y}_n = \frac{s_{xy}}{s_x^2}(x - \bar{x}_n)$$

ただし，

$$\begin{aligned} \bar{x}_n &= \frac{1}{n}\sum_{i=1}^n x_i, \qquad \bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i \\ s_x^2 &= \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x}_n)^2 \\ s_{xy} &= \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x}_n)(y_i-\bar{y}_n) \end{aligned}$$

lm：回帰直線を求める

lm と abline 身長と体重の散布図を作成

```
> height
 [1] 148 160 159 153 151 140 156 137 149 160 151 157 157 144
> weight
 [1] 41 49 45 43 42 29 49 31 47 47 42 39 48 36
> lm(weight~height)  # x 軸を身長, y 軸を体重として回帰直線を求める
Call:
lm(formula = weight ~ height)
Coefficients:
(Intercept)       height
   -70.1505       0.7399  # 切片と傾き
> bodylm<-lm(weight~height)
> plot(height,weight)  # 散布図を作成. 身長を $ x $ 軸
> abline(bodylm)       # 散布図に回帰直線を書き入れる
```

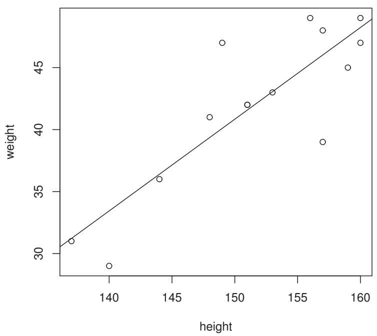

Figure 5: 散布図に回帰直線を書き入れたもの

lm：回帰直線を求める（体重を x 軸で身長を y 軸）

lm と abline 身長と体重の散布図を作成

```
>
> lm(height~weight)
Call:
lm(formula = height ~ weight)
Coefficients:
(Intercept)       weight
   110.4431       0.9792   # 切片と傾き
>
> bodylm2<-lm(height~weight)
> plot(weight,height)  # 散布図を作成．体重を x  軸
> abline(bodylm2)      # 散布図に回帰直線を書き入れる
>
```


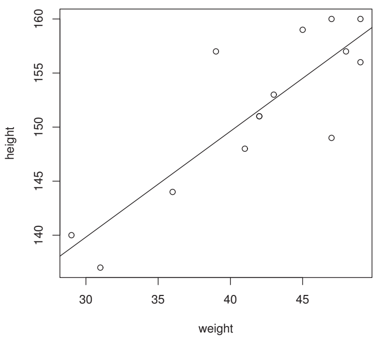

Figure 6: 散布図に回帰直線を書き入れたもの(体重が x 軸)

問題 2

(a) $Q(a, b)$ を a と b について偏微分した式を確認せよ．また，

$$\begin{cases} \frac{\partial Q}{\partial a} = 0 \\ \frac{\partial Q}{\partial b} = 0 \end{cases}$$

の解が Q を最小にする a と b の値になるのはなぜか？

(b) 連立方程式 (1) の解がつぎであることを示せ．

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2}, \qquad a = \bar{y}_n - \frac{s_{xy}}{s_x^2}\bar{x}_n$$

(c) (2) の一番右の等号を確認せよ.

- 締め切りは 2015 年 10 月 23 日（金）13 時
- このレポートは A4 のレポート用紙にかき，数研前のレポート入れに提出すること.

データより求めた回帰直線の方程式

⋆「第 I と第 III 象限にあるデータ数」＞「第 II と第 IV 象限にあるデータ数」

$\Longrightarrow$ データの全体は右上がりの傾向

⋆「第 I と第 III 象限にあるデータ数」＜「第 II と第 IV 象限にあるデータ数」

$\Longrightarrow$ データの全体は右下がりの傾向

	$x_i - \bar{x}_n$	$y_i - \bar{y}_n$	$(x_i - \bar{x}_n)(y_i - \bar{y}_n)$
I	$+$	$+$	$+$
III	$-$	$-$	$+$
II	$-$	$+$	$-$
IV	$+$	$-$	$-$

第 4 章　2 変量データの分析

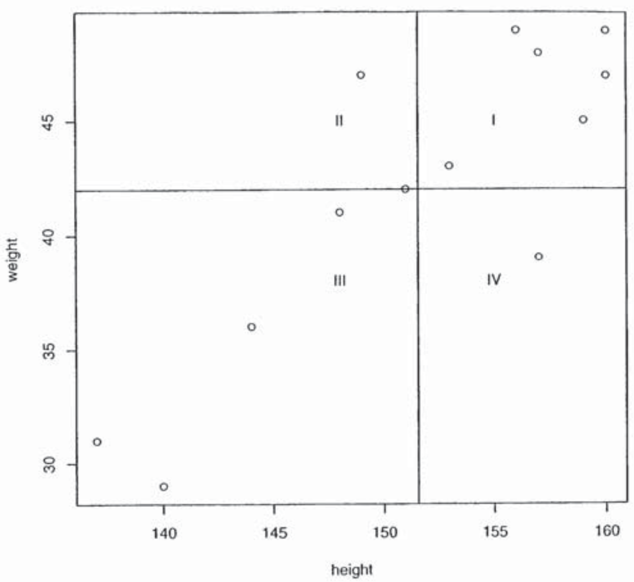

図 4.6　平均で分割した「象限」

⋆「第 I と第 III 象限にあるデータ数」＞「第 II と第 IV 象限にあるデータ数」ならば,

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) \qquad \text{の値は正}$$

⋆「第 I と第 III 象限にあるデータ数」＜「第 II と第 IV 象限にあるデータ数」ならば,

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) \qquad \text{の値は負}$$

x と y の共分散と分散

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

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2; \quad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$$

また，s_x^2 の正の平方根を s_x，s_y^2 の正の平方根を s_y と書く.

x と y の相関係数の定義

$s_x^2 > 0, s_y^2 > 0$ のとき，

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}} = \frac{s_{xy}}{s_x s_y}$$

また，s_x^2 の正の平方根を s_x，s_y^2 の正の平方根を s_y と書く.

x と y の相関係数の性質

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}} = \frac{s_{xy}}{s_x s_y}$$

- $-1 \leq r \leq 1$
- r の値が 1 に近いとき，データの全体は右上がり (正の相関が強い)
- r の値が -1 に近いとき，データの全体は右下がり (負の相関が強い)
- r の値が 0 に近いとき，無相関

相関係数を求める

cor：データ iris の散布図と相関係数

```
> x1<-iris$ Sepal.Length   # データ iris
> x2<-iris$ Sepal.Width
> x3<-iris$Petal.Length
> x4<-iris$Petal.Width
> op<-par(mfrow=c(2,2))    # 4つの散布図を同時に各コマンド
> plot(x1,x2)
> plot(x1,x3)
> plot(x1,x4)
> plot(x2,x3)
> cor(x1,x2)
[1] -0.1175698
> cor(x1,x3)
[1] 0.8717538
> cor(x1,x4)
[1] 0.8179411
> cor(x2,x3)
[1] -0.4284401
```


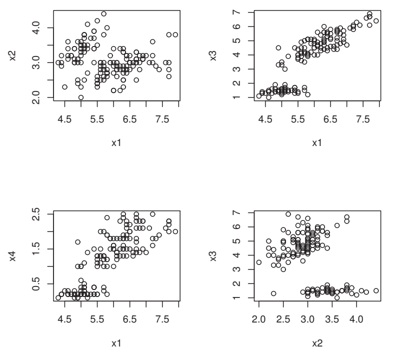

Figure 7: データ iris の散布図：相関係数　-0.1175698（右上）・0.8717538（左上）・0.8179411（右下）・-0.4284401

演習

1. 次の 2 変量データを分析せよ（散布図を用いて相関係数など）

50m 走（秒）	走り幅跳び (m)	50m 走（秒）	走り幅跳び (m)
6.8	489	7.0	398
7.2	464	7.1	485
6.8	430	7.2	400
6.8	362	6.9	511
7.2	453	7.5	430
7.0	405	7.0	487
7.0	420	7.4	470
7.1	466	7.9	380
6.8	415	6.8	460
7.1	413	7.7	398
7.4	404	7.4	415
7.2	427	6.9	470
8.0	372	7.6	450
6.8	496	7.0	500
7.6	394	7.6	410
7.0	446	6.9	500
6.6	446	7.5	400
6.6	420	6.8	505
6.8	447	7.2	522

問題 3

(a) 50m 走と走り幅跳びのデータを 10 に選び出し，それぞれをオブジェクト x と y に入力せよ.

(b) オブジェクト x と y の散布図を作成せよ.

(c) 回帰直線 $y = \hat{a} + \hat{b}x$ を求め，散布図（21416***-scatter.pdf）に描きいれよ.

(d) 相関係数を計算せよ.

- 21416***-目白花子-2015-10-23.txt

順位データ

球団	A	B
中日	3	2
広島	2	6
阪神	1	3
ヤクルト	5	5
DeNA	4	1
巨人	6	4

対象	A	B
O_1	a_1	b_1
O_2	a_2	b_2
$\vdots$	$\vdots$	$\vdots$
O_n	a_n	b_n

スピアマンの順位相関係数（1）

A, B を普通のデータとみなして，相関係数を計算してみよう.

$$
\begin{aligned}
\bar{a} &= \frac{a_1 + a_2 + \cdots + a_n}{n} = \frac{1 + 2 + \cdots + n}{n} = \frac{n(n+1)/2}{n} = \frac{n+1}{2} \\
\bar{b} &= \frac{n+1}{2} \\
s_a^2 &= \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2 - (\bar{a})^2 = \frac{a_1^2 + a_2^2 + \cdots + a_n^2}{n} - (\bar{a})^2 \\
&= \frac{1^2 + 2^2 + \cdots + n^2}{n} - \left(\frac{n+1}{2}\right)^2 = \frac{n(n+1)(2n+1)/6}{n} - \left(\frac{n+1}{2}\right)^2 \\
&= \frac{(n+1)(n-1)}{12}
\end{aligned}
$$

スピアマンの順位相関係数（2）

共分散を計算するために

$$\begin{aligned}\sum_{i=1}^{n}(a_i-b_i)^2 &= \sum_{i=1}^{n}(a_i-\bar{a}+\bar{b}-b_i)^2=\sum_{i=1}^{n}\{(a_i-\bar{a})^2-2(a_i-\bar{a})(b_i-\bar{b})+(b_i-\bar{b})^2\}\\ &= \sum_{i=1}^{n}(a_i-\bar{a})^2-2\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})+\sum_{i=1}^{n}(b_i-\bar{b})^2\\ &= ns_a^2-2s_{a,b}+ns_b^2\end{aligned}$$

よって

$$s_{a,b} = \left\{s_a^2+s_b^2-\frac{1}{n}(a_i-b_i)^2\right\}\Big/2=\frac{(n+1)(n-1)}{12}-\frac{1}{2n}\sum_{i=1}^{n}(a_i-b_i)^2$$

スピアマンの順位相関係数（3）

相関係数は

$$\begin{aligned}r_{a,b} &= \frac{s_{a,b}}{\sqrt{s_a^2 s_b^2}}\\ &= \frac{\frac{(n+1)(n-1)}{12}-\frac{1}{2n}\sum_{i=1}^{n}(a_i-b_i)^2}{\sqrt{\left(\frac{(n+1)(n-1)}{12}\right)\left(\frac{(n+1)(n-1)}{12}\right)}}\\ &= 1-\frac{6}{n(n+1)(n-1)}\sum_{i=1}^{n}(a_i-b_i)^2\end{aligned}$$

これを「**スピアマンの順位相関係数**」という.

スピアマンの順位相関係数（4）

球団	A	B	A - B	$(A-B)^2$
中日	3	2	$3-2$	1
広島	2	6	$2-6$	16
阪神	1	3	$1-3$	4
ヤクルト	5	5	0	0
DeNA	4	1	$4-1$	9
巨人	6	4	$6-4$	4
合計				34

$$r_{a,b}=1-\frac{6}{6\times7\times5}\times34=1-\frac{35}{34}=\frac{1}{35}$$

スピアマンの順位相関係数（5）

スピアマンの順位相関係数の計算

```
> a<-c(3,2,1,5,4,6)
> b<-c(2,6,3,5,1,4)
> baseball<-cbind(a,b)
> cor(a,b)
[1] 0.02857143
> cor(a,b,method="spearman")
[1] 0.02857143
```

ケンドールの順位相関係数（1）

n 個の対象 ${}_1, O_2, \ldots, O_n$ から取り出した $M = {}_nC_2$ 組の 2 個の組み合わせ $(O_i, O_j)\,(1 \leq i < j \leq n)$ に対して，大小関係をつけ，二人の大小関係が「一致する組」と「不一致の組」が現れる．M 組中，一致する組数を K，不一致の組数を L とし,「**ケンドールの順位相関係数**」を

$$\tau = \frac{K - L}{M}$$

で定義する．

ケンドールの順位相関係数（2）

	A	B	w		A	B	w
(中広)	$>$	$<$	-1	(広巨)	$<$	$>$	-1
(中阪)	$>$	$<$	-1	(阪ヤ)	$<$	$<$	$+1$
(中ヤ)	$<$	$<$	$+1$	(阪De)	$<$	$>$	-1
(中De)	$<$	$>$	-1	(阪巨)	$<$	$<$	$+1$
(中巨)	$<$	$<$	$+1$	(ヤDe)	$>$	$>$	$+1$
(広阪)	$>$	$>$	$+1$	(ヤ巨)	$<$	$>$	-1
(広ヤ)	$<$	$>$	-1	(De巨)	$<$	$<$	$+1$
(広De)	$<$	$>$	-1				
				合計			-1

$$\tau = \frac{7-8}{1} = -\frac{1}{15}$$

ケンドールの順位相関係数（3）

ケンドールの順位相関係数の計算

```
> a<-c(3,2,1,5,4,6)
> b<-c(2,6,3,5,1,4)
> baseball<-cbind(a,b)
> cor(a,b)
[1] 0.02857143
> cor(a,b,method="spearman")
[1] 0.02857143
> cor(a,b,method="kendall")
[1] -0.06666667
```


多変量のグラフ表現（1）

アヤメのデータ

```
> iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1          5.1         3.5          1.4         0.2     setosa
2          4.9         3.0          1.4         0.2     setosa
3          4.7         3.2          1.3         0.2     setosa
4          4.6         3.1          1.5         0.2     setosa
5          5.0         3.6          1.4         0.2     setosa
6          5.4         3.9          1.7         0.4     setosa
7          4.6         3.4          1.4         0.3     setosa
8          5.0         3.4          1.5         0.2     setosa
9          4.4         2.9          1.4         0.2     setosa
```

多変量のグラフ表現（2）

アヤメのデータ

```
> boxplot(iris[1:4])
> pairs(iris[1:4])
```


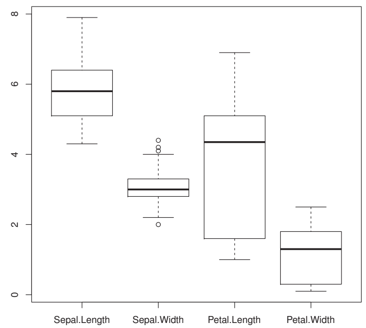

Figure 8: アイリスデータの平行箱ひげ図

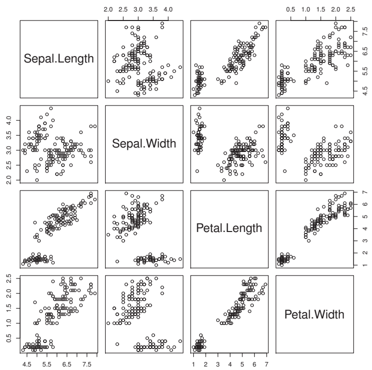

Figure 9: アイリスデータの散布図行列