確率統計と情報処理・演習（2009 年度後期）

2 変量データ分析

2009 年 10 月 23 日

日本女子大学理学部数物科学科　　今野 良彦

October 6, 2009

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」を行う.
  – ヒストグラム・箱ひげ図・分布の代表値・ばらつき
- 散布図 — 2変量のデータとして分布を眺めること.
- 回帰直線 — 散布図をみて，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 — データの直線的な傾向の強弱を数値で示すこと.

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	49	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
 [6,]    140     29
# 以下は略
> bodydata[1,]
height weight
   148     41
> bodydata[2,]
height weight
   160     49
```

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

mfrow

```
> par(mfrow=c(2,1))  # グラフを 2 行に表示
> hist(height)
> boxplot(height)
> par(mfrow=c(1,1))  # 元の戻す
```

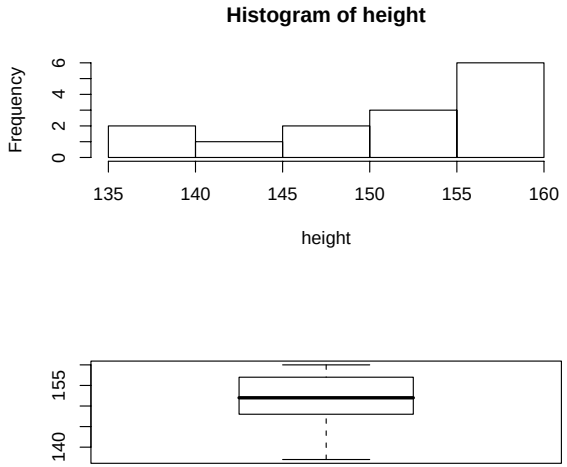

Figure 2: ヒストグラムと箱ひげ図

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それぞれ

(1) ヒストグラム
(2) 平均と中央値
(3) 箱ひげ図（mejiro-hanako-081024-boxplot.pdf）
(4) 範囲，四分位偏差，平均偏差，分散，標準偏差

を求めよ

- 締め切りは 2009 年 10 月 30 日（金）13 時
- 目白花子-091030.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
>
```

散布図

身長と体重の散布図を作成

data1.txt

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

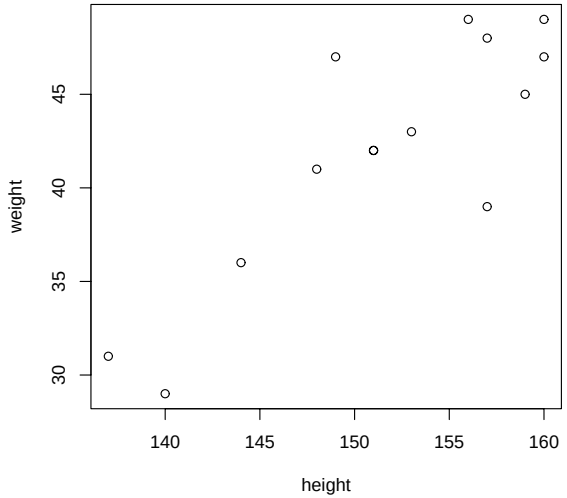

Figure 3: 身長と体重の散布図

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)
```

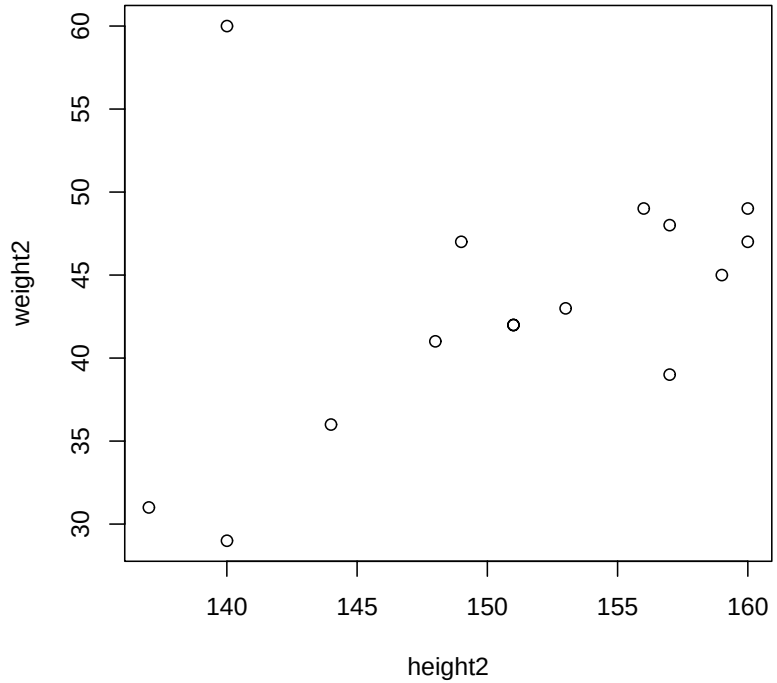

Figure 4: 身長と体重の散布図(外れ値のある場合)

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２乗法（最小自乗法）をつかおう！(他の方法もある！)

最小２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２乗法

$\star$ a と b について Q を偏微分すると

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

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

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

よって，

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

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

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

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
```

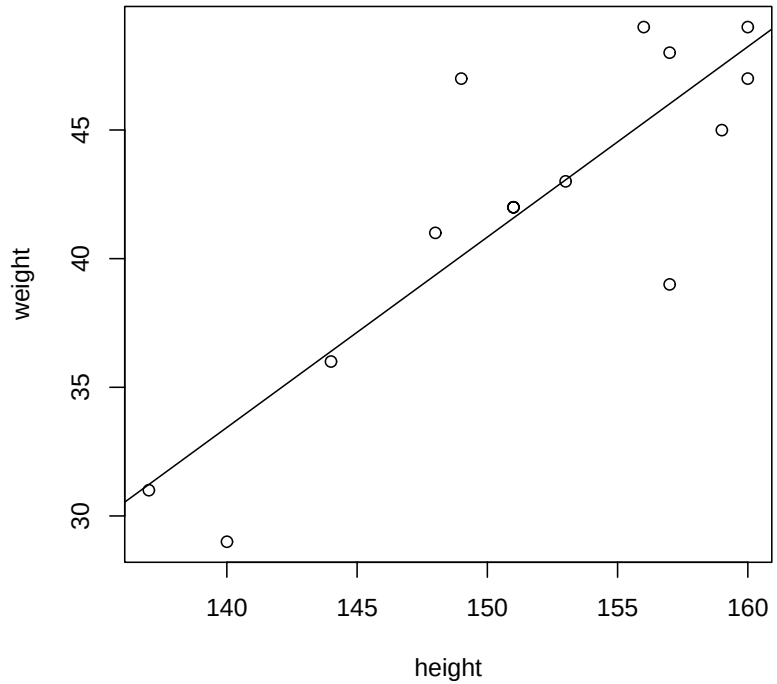

Figure 5: 散布図に回帰直線を書き入れたもの

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
>
```

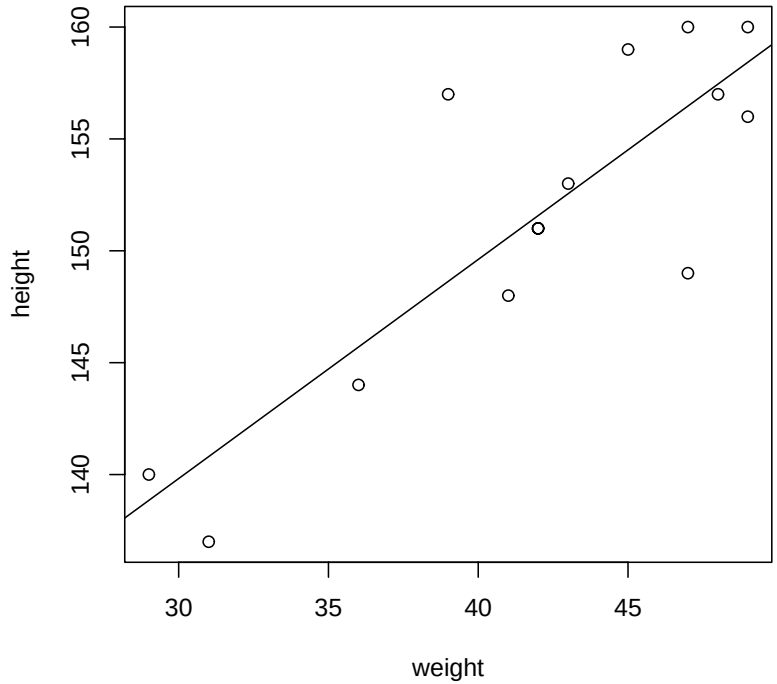

Figure 6: 散布図に回帰直線を書き入れたもの (体重が x 軸)

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，

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

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

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

(c) (2) の一番右の等号を確認せよ.

- 締め切りは 2009 年 10 月 30 日（金）13 時
- このレポートは A4 の紙にかき，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」ならば,
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」ならば,
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 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,

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

⋆「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,

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

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

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

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
```

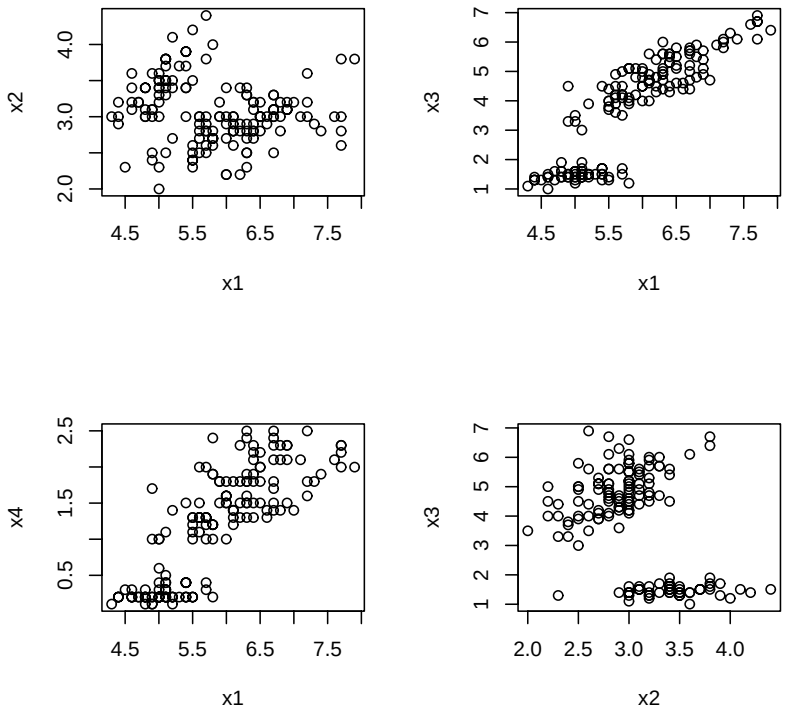

Figure 7: データ iris の散布図：相関係数　-0.1175698（右上）・0.8717538（左上）・0.8179411（右下）・-0.4284401

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(（目白花子-091030-scatter.pdf）に描きいれよ.

(d) 相関係数を計算せよ.

- 目白花子-091030.txt