2 次元データと相関

樋口さぶろお　https://hig3.net

龍谷大学理工学部数理情報学科

確率統計☆演習　L04(2020-10-19 Mon)

最終更新: Time-stamp: "2020-10-17 Sat 14:15 JST hig"

今日の目標

- データから共分散, 相関係数を計算できる

  岩薩林 確率・統計 S1.3.1
- 1 次式で平均値, 分散, 共分散, 相関係数を変換できる

  岩薩林 確率・統計 S1.3.2

L03-Q1 Quiz 解答:箱ひげ図の比較

1. $6000/8000 = 3/4$. 第 3 四分位数を答えて, 170cm.
2. $500/2000 = 1/4$. 第 1 四分位数を答えて, 190cm.
3. group A で身長が 170cm 以上 175cm 未満の人は $8000 \times 1/4 = 2000$ 人. group B で身長が 170cm 以上 190cm 未満の人は $2000 \times 1/2 = 1000$ 人. よって, 2.0 倍.

L03-Q2 Quiz 解答:箱ひげ図の比較

1. $6000/8000 = 3/4$. 第 3 四分位数を答えて, 170cm.
2. $500/2000 = 1/4$. 第 1 四分位数を答えて, 190cm.
3. group A で身長が 170cm 以上 175cm 未満の人は $8000 \times 1/4 = 2000$ 人. group B で身長が 170cm 以上 190cm 未満の人は $2000 \times 1/2 = 1000$ 人. よって, 2.0 倍.

L03-Q3 Quiz 解答:平均値・分散・標準偏差 平均値 $\bar{x} = 90\text{kg}$, 分散 $S^2 = 4\text{kg}^2$, 標準偏差 $S = 2\text{kg}$.

L03-Q4

Quiz 解答:平均値・分散・標準偏差の 1 次式による変換

1.6m, 0.0025m^2, 0.05m.

L03-Q5

Quiz 解答:分散の意味

1

L03-Q6 Quiz 解答:標準得点と偏差値 平均値 $\bar{x} = 90$, 分散 $S_x^2 = 4$, 標準偏差 $S_x = 2$.

標準得点 $z = (87 - 90)/2 = -1.5$.

偏差値 $w = (-1.5) \times 10 + 50 = 35$.

ここまで来たよ

3. データの散布度

4. 2 次元データと相関
   - 2 次元データと散布図
   - 2 次元データの相関

2 次元データ 岩薙林 確率・統計 §1.3

これまでやってたのはぜんぶ 1 次元データ.
2 次元データはこんな例. (x, y) などと書く.

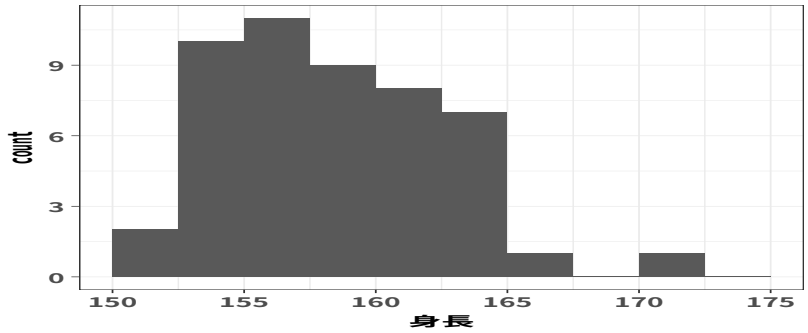
- x 身長 (cm)
- y 靴のサイズ仮 (cm) 非公表なので説明のために想像上のデータを作りました.

(メンバー)	x	y
メンバー 1	153	21.8
メンバー 2	160	24.2
⋮	⋮	⋮
メンバー 49	152	23.0
中央値	155.3	23.5
平均値	155.2	23.8
標準偏差	5.2	2.2

他にも…$(x, y) =$
(人口 (人), 面積 (m^2)),
(打率, 本塁打数),
(カロリー, 糖分含有量)....

散布図 scattered plot

岩薙林 確率・統計 §1.3(p.19)

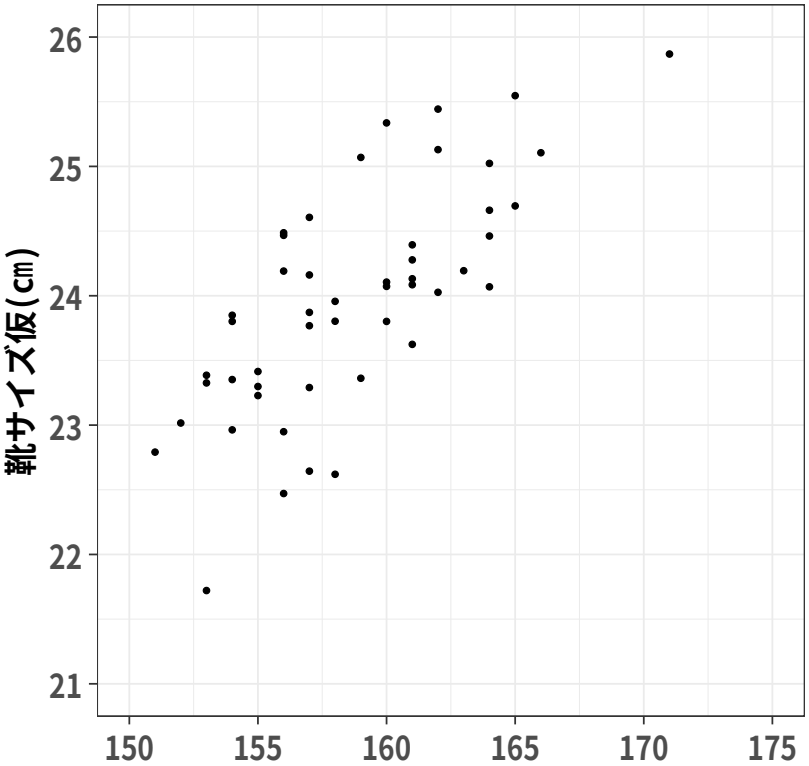


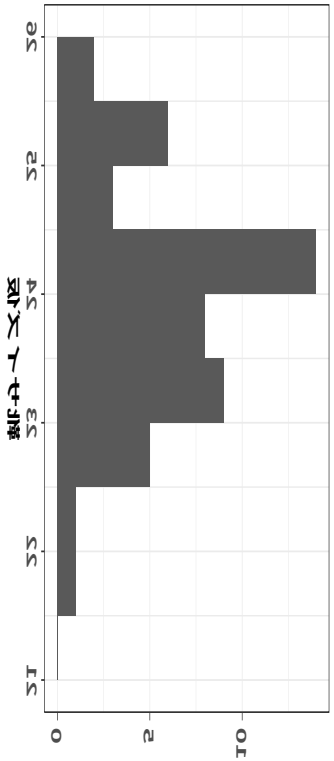
メンバー 1 人の (x, y) に点を 1 個. ストリップチャートに似てる.

重なると不便.

周辺分布

x, y の一方に着目した分布, 端, 小計





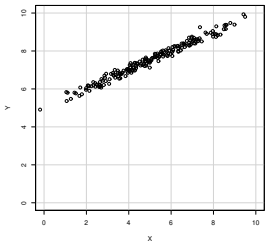
ここまで来たよ

3. データの散布度

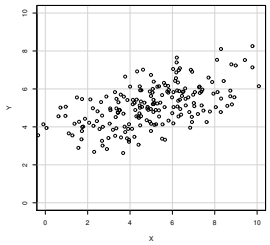
4. 2 次元データと相関
   - 2 次元データと散布図
   - 2 次元データの相関

正の相関・負の相関・無相関

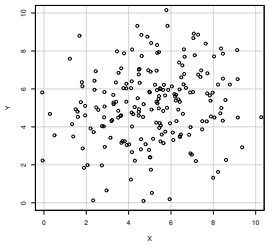
岩薩林 確率・統計 §1.3(pp.20,23)



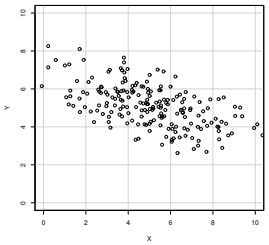
強い正の相関
$r_{xy} = 0.99$



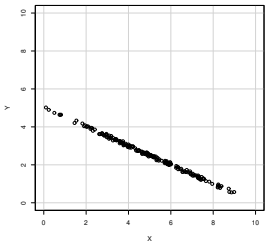
弱い正の相関
$r_{xy} = 0.55$



無相関
$r_{xy} = 0$



弱い負の相関
$r_{xy} = -0.55$



強い負の相関
$r_{xy} = -0.99$

相関

‘正の/負の相関がある’: x が大きい $\Leftrightarrow$ y が大きい/小さい傾向がある

‘相関が強い/弱い’: 傾向がはっきりしている/していない

r_{xy}: 相関係数 計算方法は以下.

共分散 高校 数学 I 発展 岩薩林 確率・統計 §1.3(p.21)

相関の正負と強さを相関係数 r_{xy} という数で表す．準備．

$$x \text{ の平均値} \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$x \text{ の分散 } S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \times (x_i - \overline{x}) = \overline{x^2} - \overline{x}^2$$

$\overline{y}, S_y^2$ も同様．

共分散 (covariance) 岩薩林 確率・統計 p.18(1.17)

$$x, y \text{ の共分散 } S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \times (y_i - \overline{y})$$

$S_{xx} = S_x^2$ みたいな感じ．

共分散公式 (便利な (こともある) 計算方法)

岩薩林 確率・統計 定理 1.5(1.18)

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y}$$

岩薩林 確率・統計 例題 1.7, 問題 7(p.19) 共分散公式での解法.

L04-Q1

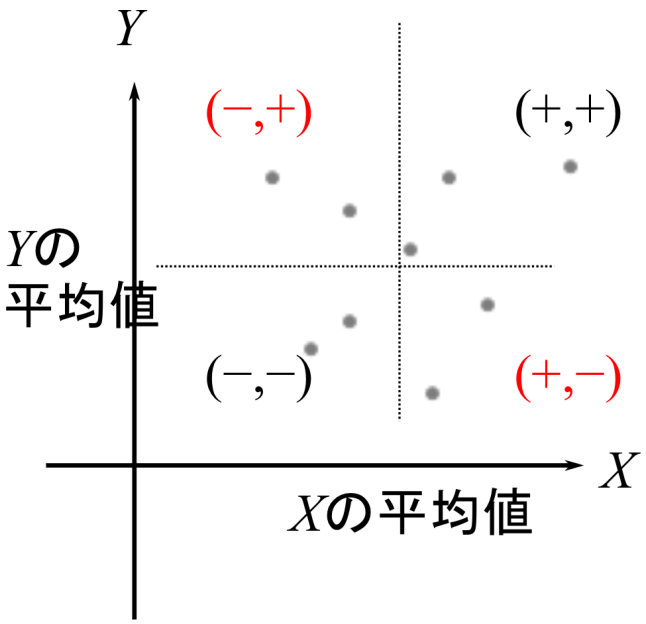
Quiz(分散共分散をモーメントから)

データ x_i, y_i $(i = 1, \ldots, n)$ に対して,
$\overline{x} = 5, \overline{y} = 10, \overline{x^2} = 34, \overline{xy} = 39, \overline{y^2} = 136$ である.

1. 分散 S_x^2 を求めよう.
2. 共分散 S_{xy} を求めよう.

共分散の意味

岩薩林 確率・統計 p.21



$(+,-) = ((x_i - \bar{x})$ の符号, $(y_i - \bar{y})$ の符号$)$.

共分散が正に/負に大きい $\Leftrightarrow$ 正の/負の相関が強い (?)

相関係数 高校 数学 I 岩薩林 確率・統計 §1.3(p.22)

共分散は

- 単位のある量. 単位を変えると 値が変わる → 比較に不便.
- 広い範囲にばらついていたほうが 大きくなる
- 1 次式の変換 $x = bu + a, y = dv + c$ で値変わる 岩薩林 確率・統計 定理 1.7

$$S_{xy} = S_{bu+a\ dv+c} = \frac{1}{n}\sum_i (bu_i+a-(b\overline{u}+a))(dv_i+c-(d\overline{v}+c)) = bdS_{uv}.$$

相関係数は, これらの影響を受けずに, 相関の強さをそのまま表す.

相関係数 (correlation coefficient) 岩薩林 確率・統計 (1.19)p.22

$$x, y \text{ の相関係数 } r_{xy} = \frac{S_{xy}}{S_x \times S_y}$$

共分散公式じゃない解法.

L04-Q2

Quiz(共分散)

1. x, y の共分散を求めよう
2. x, y の相関係数を求めよう. ただし, y の標準偏差 $= \sqrt{\frac{122}{5}} = 4.94$ は使っちゃっていい.

x	y
1	5
3	15
4	14
5	11
7	20

相関係数の性質

- $-1 \leq r_{xy} \leq +1$ 岩薩林 確率・統計 定理 1.6
- r_{xy} が正負 $\Leftrightarrow$ 正負の相関
- $|r_{xy}|$ が 0/1 に近い $\Leftrightarrow$ 相関が弱い/強い
- $r_{xy} = 0 \Leftrightarrow$ ‘相関がない’ しかし… 後ろのスライド
- $r_{xy} = \pm 1 \Leftrightarrow$ 散布図の点が傾き正/負の一直線上 $\Leftrightarrow$ y は x の 1 次式.
- r_{xy} は x, y の 1 次式による変換のもとで符号を除いて不変

$$r_{bu+a\,y} = \frac{S_{bu+a\,y}}{\sqrt{S^2_{bu+a}}\sqrt{S^2_y}} = \frac{b \cdot S_{uy}}{|b|\sqrt{S^2_u}\sqrt{S^2_y}} = \frac{b}{|b|} r_{uy} = \pm r_{uy}.$$

- 相関係数は 単位のない量

岩薩林 確率・統計 例題 1.8, 問題 8(p.22), 第 1 章練習問題 4

L04-Q3

Quiz(相関係数の性質)

2 変量データ (x, y) の相関係数を考える.

1. x に一斉に 5 を加えたとき, 相関係数はどうなる?
2. x を一斉に 2 倍したとき, 相関係数はどうなる?
3. y を一斉に -2 倍したとき, 相関係数はどうなる?
4. x, y をともに一斉に -2 倍したとき, 相関係数はどうなる?

散布図の点が傾き正/負の一直線上 $\Rightarrow r_{xy} = \pm 1$ であることの証明

$y_i = ax_i + b$ とすると.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}) \cdot ((ax_i + b) - (a\overline{x} + b)) = aS_x^2$$

ところで, $S_y = |a|S_x$ なので,

$$r = \frac{aS_x^2}{S_x|a|S_x} = \pm 1.$$

だまされたくない相関の性質

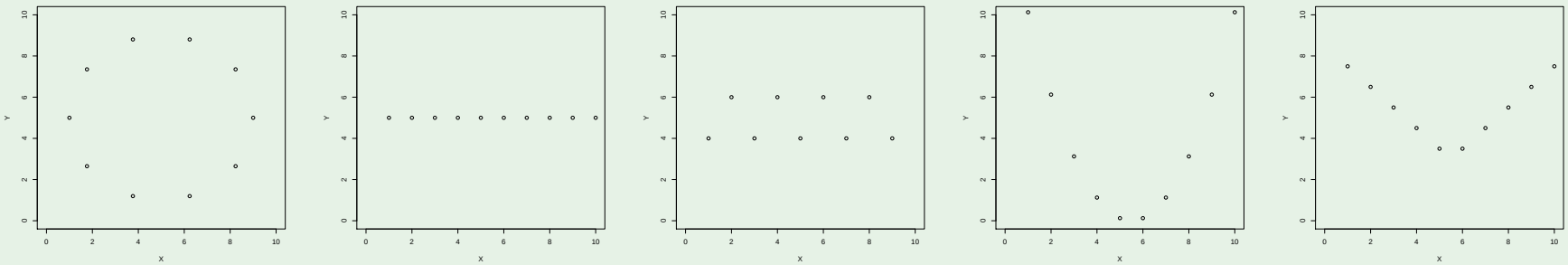
相関がある $\not\Rightarrow$ 因果関係 (原因結果の関係) がある

相関係数 $r = 0$ だから x, y は無関係な量, というわけではない

L04-Q4

Quiz(相関係数)

次のうち, 相関係数 r がもっとも大きいものはどれ?



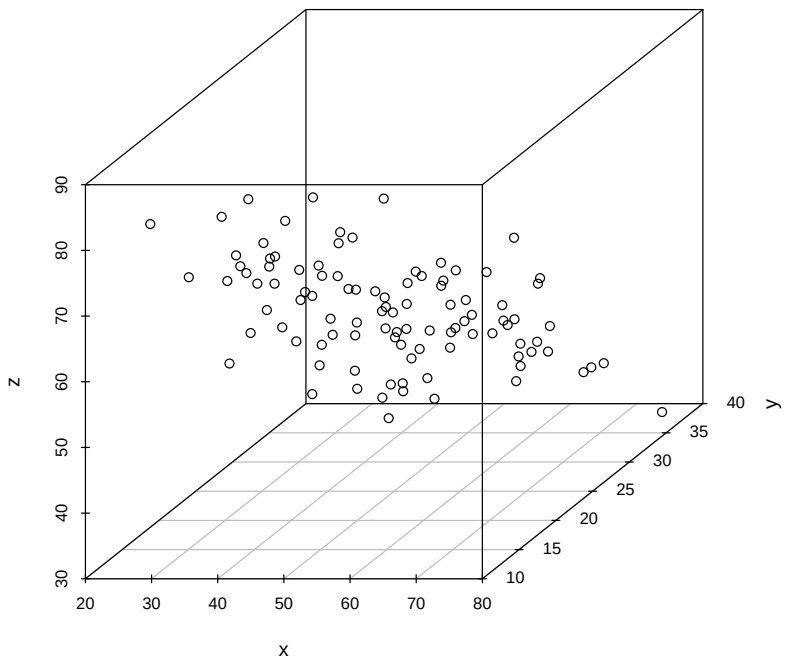
岩薩林 確率・統計 p.24

Anscombe(1973)

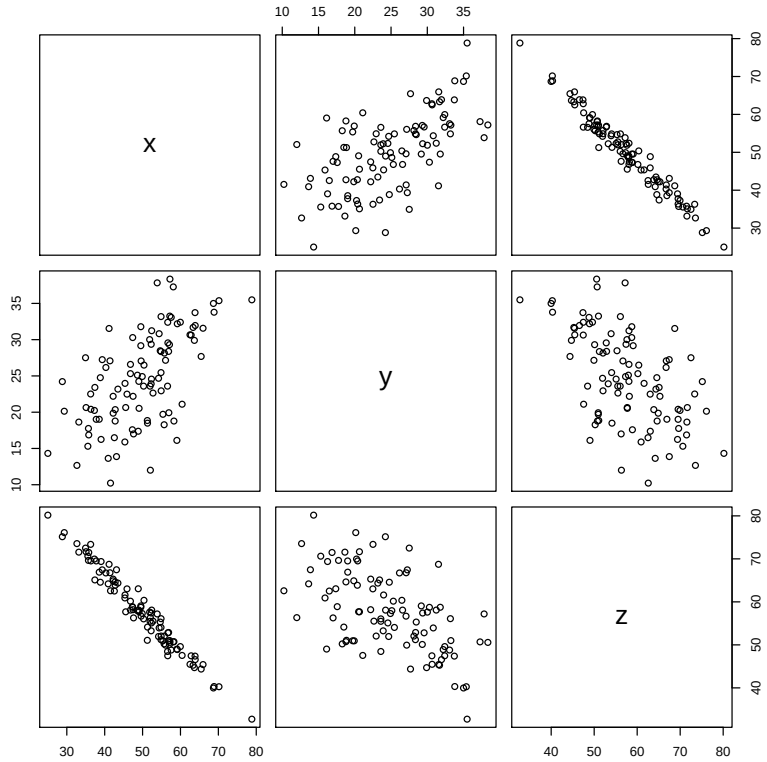
3 次元以上の場合 ($x, y \rightsquigarrow x, y, z$)

散布図

$\rightsquigarrow$ 3 次元散布図



$\rightsquigarrow$ 2 個組で散布図行列



共分散 $S_{xy} \rightsquigarrow S_{xy} = S_{yx}, S_{yz} = S_{zy}, S_{zx} = S_{xz}$. 2 個組で.
共分散行列 (対称行列)

$$\begin{bmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{bmatrix} \rightsquigarrow \begin{bmatrix} S_x^2 & S_{xy} & S_{xz} \\ S_{xy} & S_y^2 & S_{yz} \\ S_{xz} & S_{yz} & S_z^2 \end{bmatrix}$$

相関係数 $r_{xy} \rightsquigarrow r_{xy} = r_{yx}, r_{yz} = r_{zy}, r_{zx} = r_{xz}$. 2 個組で.
相関行列 (対称で, 対角成分が $r_{ii} = 1$ で, 全成分 $|r_{ij} \leq 1|$)

$$\begin{bmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{bmatrix}$$