データの変換 (標準得点, 偏差値)・2 変量データと相関

樋口さぶろお

龍谷大学理工学部数理情報学科

確率統計☆演習 I L03(2018-10-10 Wed)

最終更新: Time-stamp: "2018-10-10 Wed 07:06 JST hig"

今日の目標

- データを 1 次関数で変換したときの平均値, 分散が求められる
- データの標準得点, 偏差値を求められる
- 2 変量データの共分散, 相関係数が求められる

L02-Q1

Quiz 解答:代表値

1. $Q_2 = 17\text{cm}, Q_1 = 14.5\text{cm}, Q_3 = 18\text{cm}$.
2. 最頻値は 18cm.
3. 平均値は $(14 + \cdots + 25)/8 = 17.25\text{cm}$.

L02-Q2

Quiz 解答:平均値中央値最頻値 $N = 9$.

1. 中央値 $Q_2 = x_4$. よって階級 21–23 に含まれる. 近似値として階級値を答えて $x_4 \approx 21 + 2 \times \frac{1.5}{2} = 22.5$.
2. 階級値を答えて, 10
3. 階級値で近似して, $\frac{1}{9}(10 \times 3 + 22 \times 2 + 24 \times 2 + 26 \times 2) = 19.3$

L02-Q3

Quiz 解答:範囲

範囲は $Q_4 - Q_0 = 25 - 14 = 11$, 四分位範囲は $Q_3 - Q_1 = 18 - 14.5 = 3.5$, 四分位偏差は $\frac{1}{2}(Q_3 - Q_1) = 1.75$.

L02-Q4 Quiz 解答:平均値・分散・標準偏差 平均値 $= 90\text{kg}$, 分散 $= 4\text{kg}^2$, 標準偏差 $= 2\text{kg}$.

ここまで来たよ

平均値・分散・標準偏差の変換

x から y への変換

データ $x_1, x_2, \ldots, x_n$, x の平均値 $\overline{x}$, 分散 S_x^2, 標準偏差 S_x がわかってるとする.

$y_i = ax_i + b$ で新しいデータを作る (a, b 定数).

データ $y_1, y_2, \ldots, y_n$, y の平均値 $\overline{y}$, 分散 S_y^2, 標準偏差 S_y はどうやって求める?

例: 身長の換算 $y = 1.8(\mathrm{m}) \leftarrow x = 80(\mathrm{cm})$

$y = ax + b$, $\boxed{}$

平均値, 分散, 標準偏差の変換

$y = ax + b$ のとき

1. $\overline{y} = a\overline{x} + b$
2. $S_y^2 = |a|^2 \times S_x^2$
3. $S_y = |a| \times S_x$

L03-Q1

Quiz(平均値・分散・標準偏差の換算)

ある集団の身長 (みんな大人で 100cm 以上) を, cm で書いたものの下 2 桁 x cm の, 平均値は 60cm, 分散は 25cm^2 だった.
m で書いた身長 y m の平均値と分散と標準偏差を求めよう.

ここまで来たよ

標準偏差の意味 I

L03-Q2

Quiz(分散の意味)

あるクラスで行われたテストで, 英語の平均点は 60 点, 標準偏差 10 点.
数学の平均点は 60 点, 標準偏差 20 点.
英語の 70 点と数学の 70 点, どちらのほうが価値ある (上位にいる可能性が高い)? 次のうちから正しいものを 1 つ選ぼう.

1. たぶん英語のほうが価値ある
2. たぶん数学のほうが価値ある
3. どちらも同じ
4. 追加の情報がないとわからない
5. 追加の情報があっても比べることはできない

標準得点 I

標準得点 (standard score, z-score, z 得点)

$$(\text{値 } x_i \text{ の}) \text{ 標準得点 } z_i = \frac{x_i - \overline{x}}{S_x}$$

平均値から, 上下どちらに, 標準偏差の何倍離れているかを表す値.

例 $N = 5$

i	1	2	3	4	5	平均値	標準偏差
データ x_i	15	13	12	11	9	12	2
標準得点 z_i	1.50	0.5	0	-0.5	-1.50	0	1

L03-Q3

Quiz(標準得点と偏差値)

データ 87, 93, 89, 91, 90 で, 87 の標準得点と偏差値を求めよう.

標準得点の性質

標準得点 z の性質

- $\overline{z} =$ □
- $S_z^2 =$ □, $S_z =$ □
- z の単位は □, 無次元の数. 身長が 180cm, 80cm, 1.8m どれでも同じ結果.

なぜなら… いま □.

$$\overline{z} = a\overline{x} + b = \frac{1}{S_x} \cdot \overline{x} - \frac{\overline{x}}{S_x} = 0.$$

$$S_z = |a| S_z = \left| \frac{1}{S_x} \right| S_x = 1.$$

偏差値

学力データ (テストの点数や成績?) によく使われる.
受験者 1 人 1 人の成績が, 平均値から上, または下に離れている程度を見られる.

偏差値

$$\begin{aligned}(\text{値 } x_i \text{ の}) \text{ 偏差値 } w =& 10z_i + 50 \\ =& \frac{x_i - \overline{x}}{S_x} \times 10 + 50.\end{aligned}$$

$a =$ [　　], $b =$ [　　]

- 異なるテストでも比べられる.
- 偏差値の平均値は [　], 偏差値の標準偏差は [　]
- 偏差値はまあ ‘無次元の数’(1000 点満点と 100 点満点を比較可能)

L03-Q4

Quiz(偏差値の性質)

次を, 正しい, 誤り, もっともらしいが正しいとは断定できない, に分類しよう.

1. 別の塾に転校した後, 塾内テストの偏差値が上がったことから, 成績が上がったと言える.
2. 同じ学級内の偏差値が, 中間試験より期末試験で下がったので, 学級内の順位が下がったと言える.
3. 教員が全受験者に 5 点を加点したので, 偏差値は実際より高めに出ているはずである.
4. 同じ学級内での偏差値が, 数学より理科のほうが高いので, 理科のほうがより上位にいると言える.

ここまで来たよ

2. 略解:データの代表値と散布度

3. データの変換 (標準得点, 偏差値)・2 変量データと相関
   - 分散の意味と平均値・分散・標準偏差の変換
   - 標準得点・偏差値

4. 2 変量データの相関
   - 2 変量データと散布図
   - 2 変量データの相関

2 変量データ 前園確率統計 §4.3

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

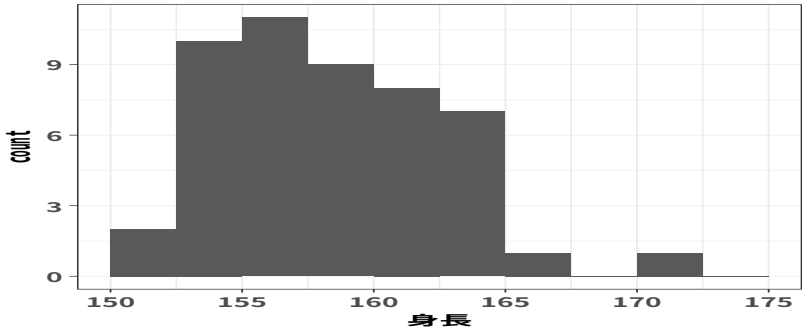
- x 身長 (cm)
- y 靴のサイズ仮 (cm) 非公表なので説明のために想像上のデータを作りました.

(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

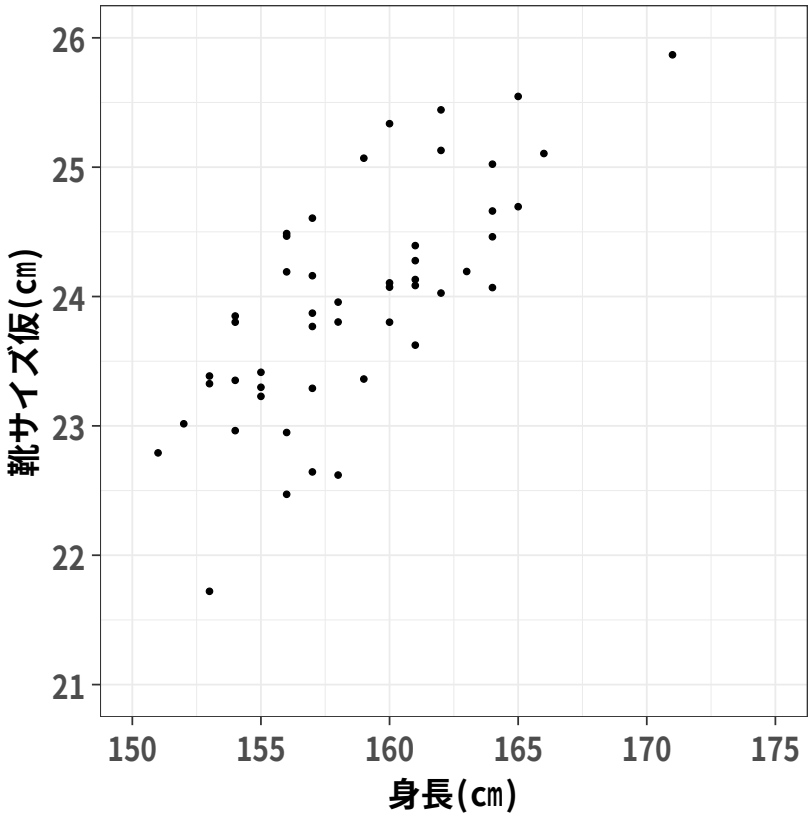
他にも…(x, y) =(人口 (人), 面積 (m^2), (打率, 本塁打数), (カロリー, 糖分含有量)….

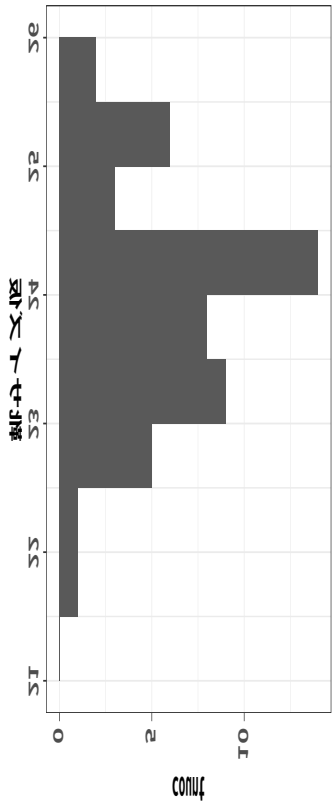
散布図=相関図

前園確率統計 §4.3



メンバー 1 人の (x, y) に点を 1 個.
不便な点は
周辺分布とは

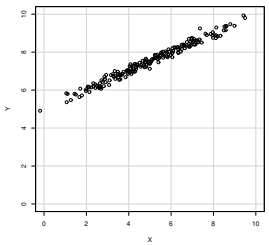




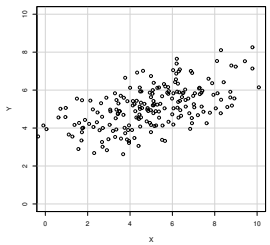
ここまで来たよ

正の相関・負の相関・無相関

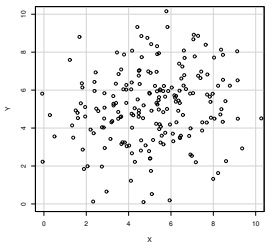
前園確率統計 §4.3



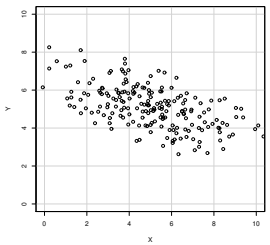
強い正の相関
$r = 0.99$



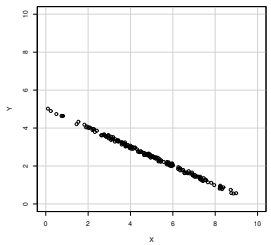
弱い正の相関
$r = 0.55$



無相関
$r = 0$



弱い負の相関
$r = -0.55$



強い負の相関
$r = -0.99$

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: 相関係数 計算方法は以下.

共分散 高校 数学 I 発展

相関の強さを相関係数 r という数で表す．復習と準備

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

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

$\overline{y}, S_y^2$ も同様．

共分散 (covariance) 前園確率統計 §4.3

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

注: $C_{xy} = S_{xy}$, x の分散を $S_x^2 = S_{xx}$, y の分散を $S_y^2 = S_{yy}$ と書く自然な記法がある．

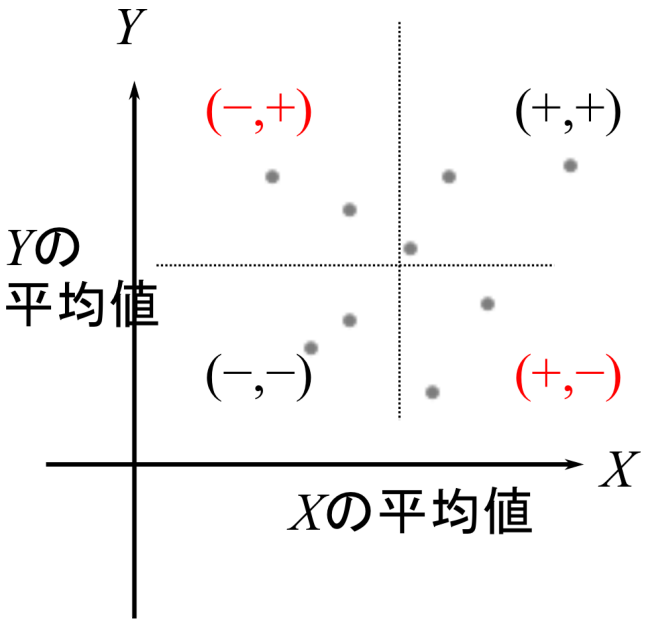
L03-Q5

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

共分散の意味



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

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

なぜなら

しか〜し (次のスライド)

相関係数 高校 数学 I

共分散は

- x, y の 1 次関数による変換で変わる

$$C_{ax+b\,y} = aC_{xy}.$$

- 単位を変えると [　　] → 比較に不便
- 広い範囲にばらついていたほうが [　　]

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)

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

相関係数の性質

- $-1 \leq r \leq +1$
- r が正負 $\Leftrightarrow$ 正負の相関
- $|r|$ が $0/1$ に近い $\Leftrightarrow$ 相関が弱い/強い
- $r = 0 \Leftrightarrow$ ' 相関がない' しかし…
- $r = \pm 1 \Leftrightarrow$ 散布図の点が傾き正/負の一直線上 $\Leftrightarrow$ y は x の 1 次関数.
- r は x, y の 1 次関数による変換のもとで符号を除いて不変

$$r_{ax+b \quad y} = \pm r_{xy}$$

.

- 相関係数は

L03-Q6

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.

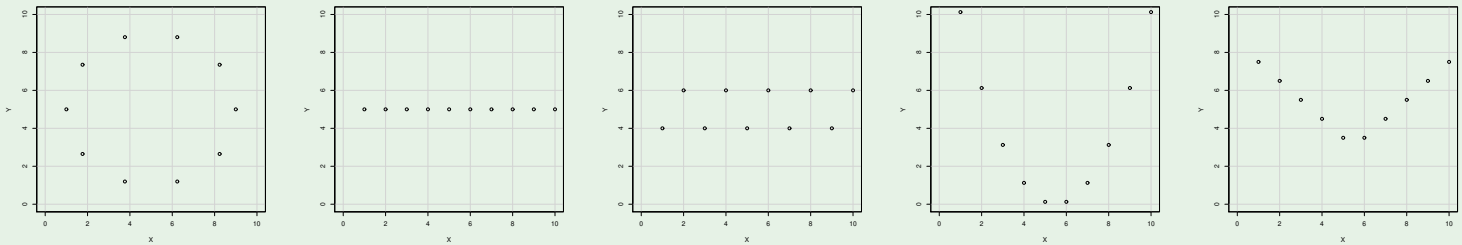
1. x に一斉に 5 を加えたとき, 相関係数はどうなる?
2. x を一斉に 2 倍したとき, 相関係数はどうなる?
3. y を一斉に -2 倍したとき, 相関係数はどうなる?
4. x, y をともに一斉に -2 倍したとき, 相関係数はどうなる?

だまされたくない相関の性質

L03-Q7

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?



Anscombe(1973)

連絡

- 次回は臨時教室変更で 1-609 実習室
- 動画見るので PC につながるイヤフォン持ってきて (Bluetooth や Lightning じゃなく)
- 樋口オフィスアワー火昼 (1-539) 金 14:40-15:40(1-502), Math ラウンジ月-木昼 (1-614)
- Trial 予告
- Learn Math Moodle の予習復習問題で来週の trial に備えてね.
- 来週は, ちょっと, 無理してるけど教科書 前園確率統計 §7.2 読んできて