データの変換 (標準得点, 偏差値)・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 I　L03(2019-10-07 Mon)

最終更新: Time-stamp: "2019-10-07 Mon 07:21 JST hig"

今日の目標

- データを 1 次式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 次元データの共分散, 相関係数が求められる

L02-Q1

Quiz 解答:平均値中央値最頻値

$n = 9.$

1. 中央値 $Q_2 = x_{\frac{1}{2}\cdot(9-1)} = x_4$. (0 番目から数え始めて 4 番目). 高校のりでいったら, 小さい順に並べた 9 個の真ん中の値. よって階級 $21 - -23$ に含まれる. 階級値 22 で近似できる.
2. もっとも度数の多い階級の階級値で, 10.
3. 階級値で近似して計算すると,
   $\frac{1}{9}(10 \times 3 + 22 \times 2 + 24 \times 2 + 26 \times 2) = 19.3.$

L02-Q2

Quiz 解答:範囲

$n = 8$ で, $x_0 = 14, \cdots, x_7 = 25$ とする.

$Q_0 = x_0 = 14$. $Q_4 = x_7 = 25$.

$Q_2 = x_{\frac{1}{2}\cdot 7} = x_{3.5}$ で $x_3 = 16$ と $x_4 = 18$ の間. 高校数学では,

$\frac{1}{2}(x_3 + x_4) = 17$.

$Q_1 = x_{\frac{1}{4}\cdot 7}$, $Q_3 = x_{\frac{3}{4}\cdot 7}$ だが, 高校数学では, $Q_1 =$

$(14, 14, 15, 16 \text{ の中間値}) = 14.5, Q_3 = (18, 18, 18, 25 \text{ の中間値}) = 18.$

範囲は $Q_4 - Q_0 = x_7 - x_0 = 25 - 14 = 11$.

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-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 $S = 2\text{kg}$.

ここまで来たよ

平均値・分散・標準偏差の変換

岩薙林 確率・統計 §1.2.2

x から y への変換

データ $u_1, u_2, \ldots, u_n$, u の平均値 $\overline{u}$, 分散 S_u^2, 標準偏差 S_u がわかってるとする.
$x_i = bu_i + a$ で新しいデータを作る (a, b 定数).
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 はどうやって求める?

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

$x = bu + a$, $\boxed{}$

平均値・分散・標準偏差の 1 次式による変換

岩薙林 確率・統計定理 1.3

$x = bu + a$ のとき

1. $\overline{x} = b\overline{u} + a$ 岩薙林 確率・統計 (1.9)
2. $S_x^2 = b^2 \times S_u^2$ 岩薙林 確率・統計 (1.10)
3. $S_x = |b| \times S_u$ 岩薙林 確率・統計 (1.11)

岩薙林 確率・統計例題 1.5(p.13), 問題 3(p.14)

L03-Q1

Quiz(平均値・分散・標準偏差の 1 次式による変換)

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 u cm の, 平均値は 60cm, 分散は 25cm^2 だった.
m で書いた身長 x 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

標準得点 (standard score, z-score, z 得点) 岩薩林 確率・統計 (1.13) 例 4(p.14)

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

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.

$u = z$ と思うと,

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 岩薩林 確率・統計問題 4(p.15)

- $\overline{z} =$ □
- $S_z^2 =$ □, $S_z =$ □
- 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 岩薩林 確率・統計問題 1.4,

$$\overline{x} = b\overline{z} + a$$
$$S_x = |b| S_z.$$

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

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

$a =$ [　　], $b =$ [　　]

- 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)

ここまで来たよ

2 データの代表値と散布度
- 平均値・分散・標準偏差の 1 次式による変換
- 標準得点・偏差値

3 2 次元データの相関
- 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.

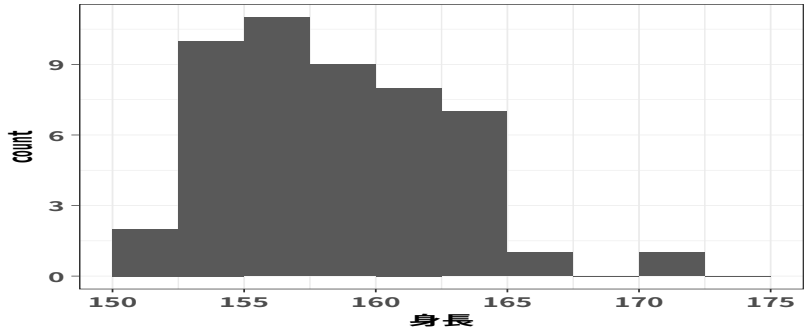
- x 身長 (cm)
- y 靴のサイズ仮 (cm) 非公表なので説明のために想像上のデータを作りました.

(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

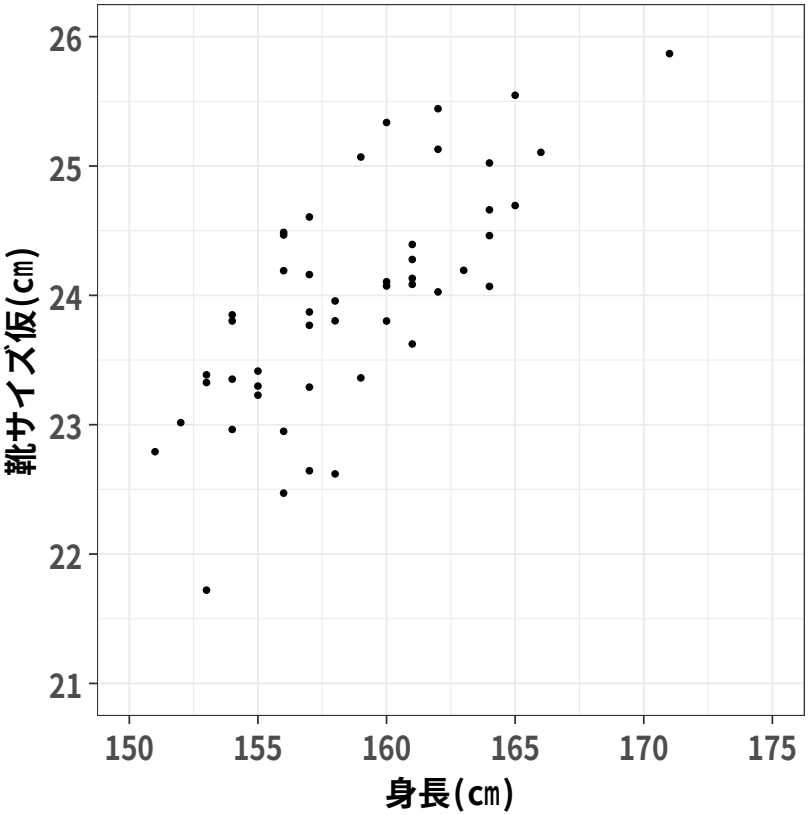
他にも…(x, y) =(人口 (人), 面積 (m^2), (打率, 本塁打数), (カロリー, 糖分含有量)….

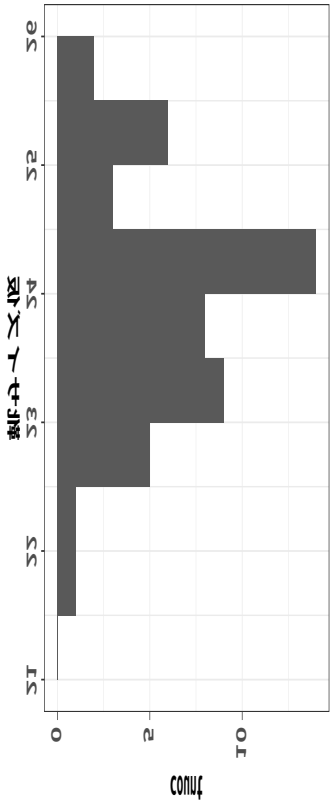
散布図

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.
不便な点は
周辺分布とは





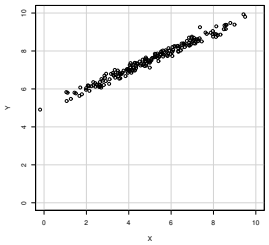
ここまで来たよ

2 データの代表値と散布度
- 平均値・分散・標準偏差の 1 次式による変換
- 標準得点・偏差値

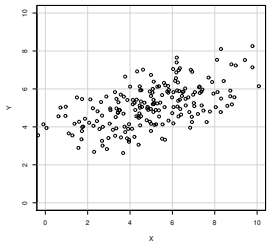
3 2 次元データの相関
- 2 次元データと散布図
- 2 次元データの相関

正の相関・負の相関・無相関

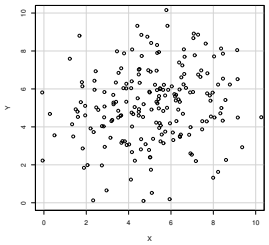
岩薙林 確率・統計 §1.3(pp.20,23)



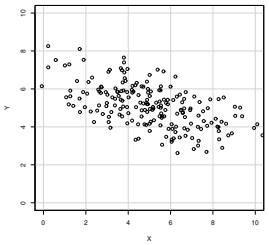
強い正の相関
$r = 0.99$



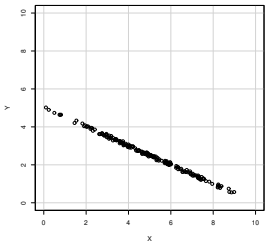
弱い正の相関
$r = 0.55$



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 岩薩林 確率・統計 §1.3(p.21)

相関の強さを相関係数 r という数で表す．復習と準備

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

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

$\overline{y}, S_y^2$ も同様．

共分散 (covariance) 岩薩林 確率・統計 p.18

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

$S_{xx} = S_x^2$ みたいな感じ．

岩薩林 確率・統計例題 1.7, 問題 7(p.19)

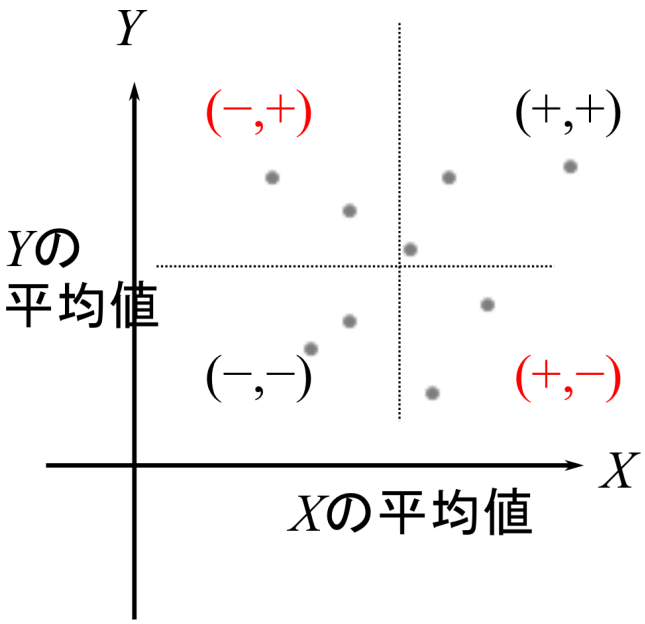
L03-Q4

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 岩薩林 確率・統計 p.21



$(+,-) = ((x_i - \overline{x})$ 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 岩薩林 確率・統計 §1.3(p.22)

共分散は

- x, y の 1 次関数による変換で変わる

$$S_{bu+a\ dv+c} = bdS_{uv}.$$

岩薩林 確率・統計定理 1.7

- 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) 岩薩林 確率・統計 (1.19)p.22

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

相関係数の性質

- $-1 \leq r \leq +1$ 岩薩林 確率・統計定理 1.6
- 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

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

.

- 相関係数は

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

L03-Q5

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.

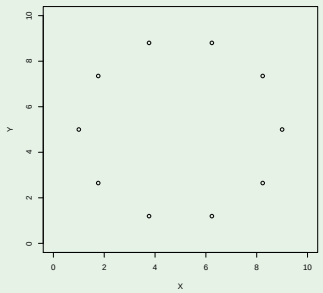
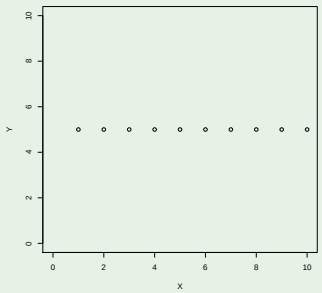
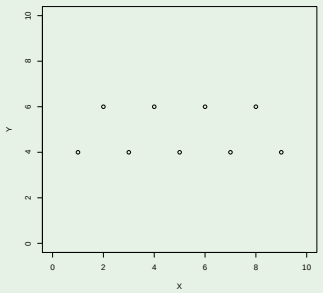
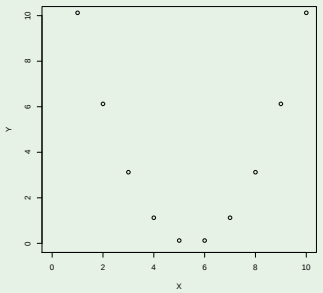
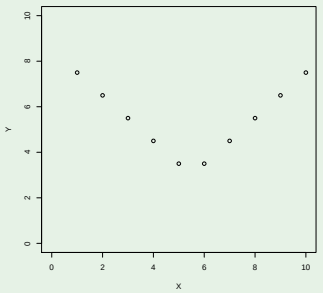
1. x に一斉に 5 を加えたとき, 相関係数はどうなる?
2. x を一斉に 2 倍したとき, 相関係数はどうなる?
3. y を一斉に -2 倍したとき, 相関係数はどうなる?
4. x, y をともに一斉に -2 倍したとき, 相関係数はどうなる?

だまされたくない相関の性質

L03-Q6

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

- 次回はたぶん 1-609 実習室
- オフィスアワー木 6(1-539) 金昼 (1-542), Math ラウンジ (1-536/538)
- Trial 予告
- 来週は教科書 岩薩林 確率・統計 §9 読んできて.

統計検定. 2019-11-24 日 40%ディスカウント団体受験受付中. 今日 2019-10-07 月まで.

Moodle モバイルアプリ で URL 指定

https://note.math.ryukoku.ac.jp/moodle