2 変量データ

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 L04(2015-10-09 Fri)

最終更新: Time-stamp: "2015-10-09 Fri 13:21 JST hig"

今日の目標

- データから標準得点, 偏差値が計算できる
- 2 変量データから共分散, 相関係数が計算できる
- 2 変量データの相関係数の意味が説明できる

http://hig3.net

L03-Q1

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 $\frac{1}{2}(Q_3 - Q_1) = 1.75$.

L03-Q2 Quiz 解答:平均値・分散・標準偏差　　平均値 $= 90$kg, 分散 $= 4\text{kg}^2$, 標準偏差 $= 2$kg.

L03-Q3 Quiz 解答:度数分布表から分散　　平均値 $= 160$(cm), 分散 $= (10^2 \times 20 + 0^2 \times 40 + 10^2 \times 20)/80 = 50$ (cm^2).

L03-Q6

Quiz 解答:箱ひげ図

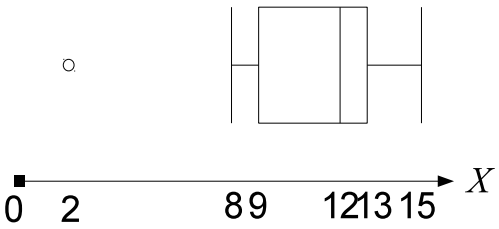
$\text{Q}_2 = 12\text{g}$.

$\text{Q}_1 = \frac{1}{2}[8 + 10] = 9\text{g}$.

$\text{Q}_3 = \frac{1}{2}[12 + 14] = 13\text{g}$.

四分位範囲 $= 13 - 9 = 4\text{g}$.

$x = 2$ は, $\text{Q}_1 = 9$ から $4 \times 1.5 = 6$ 以上大きく離れているので, 外れ値である.


0
2
8
9
12
13
13
15
X

ここまで来たよ

3 データのばらつきを表す値

4 2 変量データ
- 標準得点と偏差値
- 2 変量データとクロス集計表・散布図
- 2 変量データの相関

(復習) 平均値, 分散, 標準偏差の換算

$y = ax + b$ のとき

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

L04-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.

標準得点

標準得点 (standard score)

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

- 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.
- z-得点 (z-score) などともいう.

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

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

0–100 の範囲の値をとるデータ (テストの点数や成績?) に使われる.
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, クラスでも比べられる.
- 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)

L04-Q2

Quiz(偏差値)

(学力) 偏差値について, 次のうち正しいのはどれ (とどれ)?

1. 偏差値の最低値は 0 である
2. 偏差値の最高値は 75 である
3. 平均点 (をとった人) の偏差値は 50 である
4. 100 点のテストで満点を取った場合の偏差値は, 他の人の成績しだいである
5. 偏差値 50 の人の順位は上から 1/2 程度である
6. 偏差値 60 の人の順位は上から 15% 程度である.

L04-Q3

Quiz(標準得点と偏差値)

データ x は 87, 93, 89, 91, 90 で与えられる. 87 の標準得点と偏差値を求めよう.

ここまで来たよ

3 データのばらつきを表す値

4 2 変量データ
- 標準得点と偏差値
- 2 変量データとクロス集計表・散布図
- 2 変量データの相関

2 変量データ

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, y は各チームのデータ.

- x 勝利数
- y (打った) シュート数
- z 失点

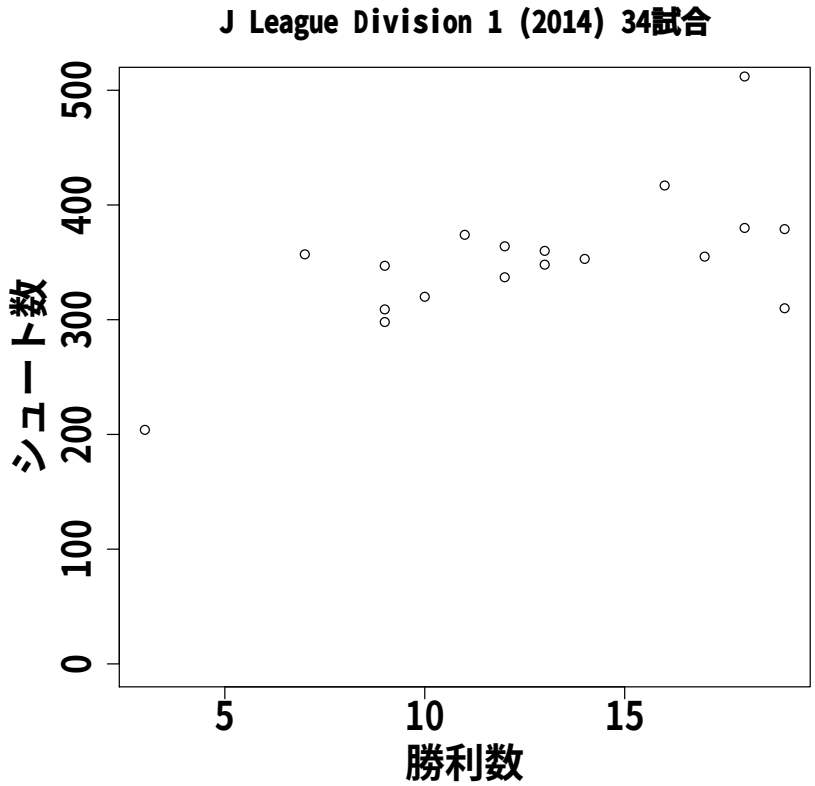
J リーグ Div1. 2014 年の 34 試合. データの個数 $n = 18$(チーム).

(チーム名)	x	y	z
ベガルタ仙台	9	347	50
鹿島アントラーズ	18	512	39
⋮	⋮	⋮	⋮
計	...	...	...
平均値	...	...	...

http://www.j-league.or.jp/data/

他にも…(x, y) =(身長 (cm), 体重 (kg)), (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


J League Division 1 (2014) 34試合
シュート数
0
100
200
300
400
500
5
10
15
勝利数


?

クロス集計表と周辺分布

x:勝利数, y (打った) シュート数

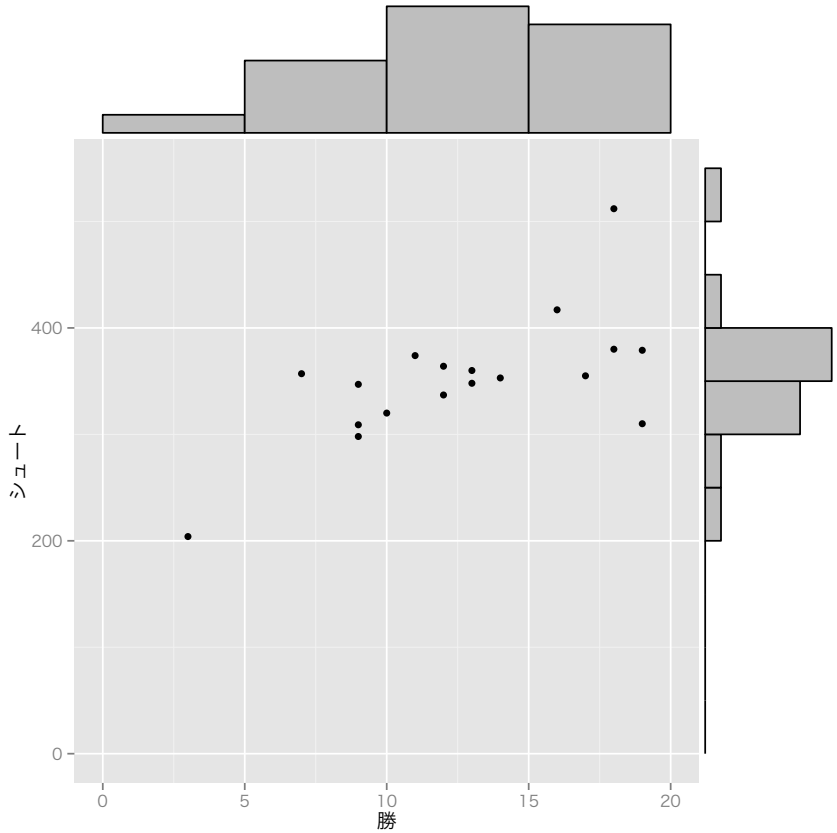
クロス集計表 度数分布表の 2 変数版

上の表では…になってる 18 チーム全部のデータから作りました.

$\downarrow y$　$\backslash x$ の階級 $\rightarrow$	0 以上 5 未満	10 未満	15 未満	20 未満	計
200 以上 250 未満	1				1
250 以上 300 未満		1			1
300 以上 350 未満		2	3	1	6
350 以上 400 未満		1	4	3	8
400 以上 450 未満				1	1
450 以上 500 未満				0	0
500 以上 550 未満				1	1
計	1	4	7	6	18

周辺分布とは

周辺分布のヒストグラム



周辺分布のヒストグラムは, 散布図で

[　　] して作れる.

L04-Q4

Quiz(クロス集計表)

1. 散布図を描こう.
2. クロス集計表を作ろう. x の階級は 0 以上 2 未満, …, y の階級は 0 以上 5 未満, … で.

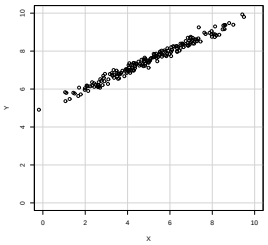
x	y
1	5
3	15
4	14
5	11
7	20

ここまで来たよ

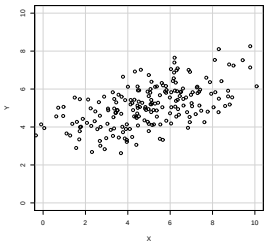
3 データのばらつきを表す値

4 2 変量データ
- 標準得点と偏差値
- 2 変量データとクロス集計表・散布図
- 2 変量データの相関

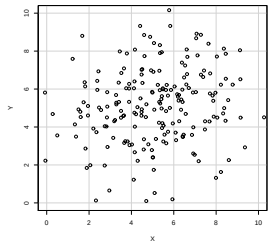
正の相関・負の相関・無相関



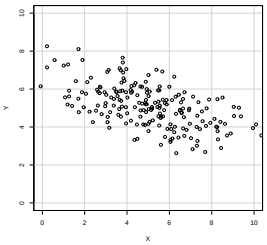
強い正の相関
$r = 0.99$



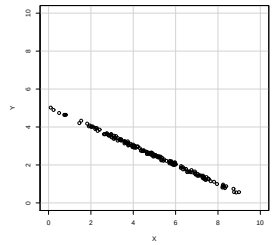
弱い正の相関
$r = 0.55$



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

‘正の相関’: x が大きい $\Leftrightarrow$ y が大きい

‘負の相関’: x が大きい $\Leftrightarrow$ y が小さい

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

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

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

$\overline{y}, s_y^2$ も同様.

共分散 (covariance)

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

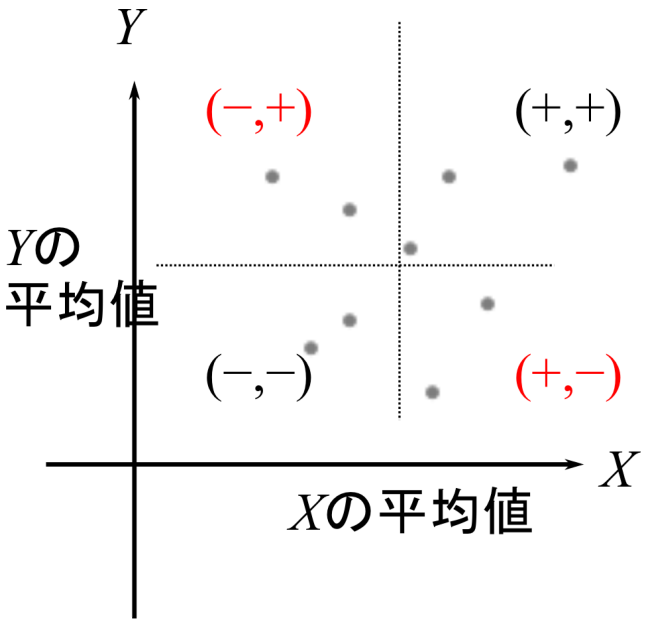
L04-Q5

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



$(+,-) = (x_i - \bar{x}$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 (?)

なぜなら

しか〜し.

相関係数

共分散は

- 次元のある量なので単位を変えると［　　］→ 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

- 相関係数は
- $-1 \leq r \leq +1$
- $r = 0 \Leftrightarrow$ ' 無相関'
- $r = \pm 1 \Leftrightarrow$ 散布図の点が傾き正/負の一直線上 $\Leftrightarrow$ y は x の 1 次関数.

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

$y_i = ax_i + b$ とすると.

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot ((ax_i + b) - (a\overline{x} + b)) = as_x^2$$

ところで, $s_y = |a|s_x$ なので,

$$r = \frac{as_x^2}{s_x|a|s_x} = \pm 1$$

.

L04-Q6 重複した問題

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

L04-Q7

Quiz(共分散と相関係数)

下の 2 変量データ (x, y) を考える.

x(cm)	y(g)
13	2
16	4
18	2
18	4
21	4
22	8

次の量を求めよう.

1. 共分散 C_{xy}
2. 相関係数 r

相関係数=0 にだまされるな

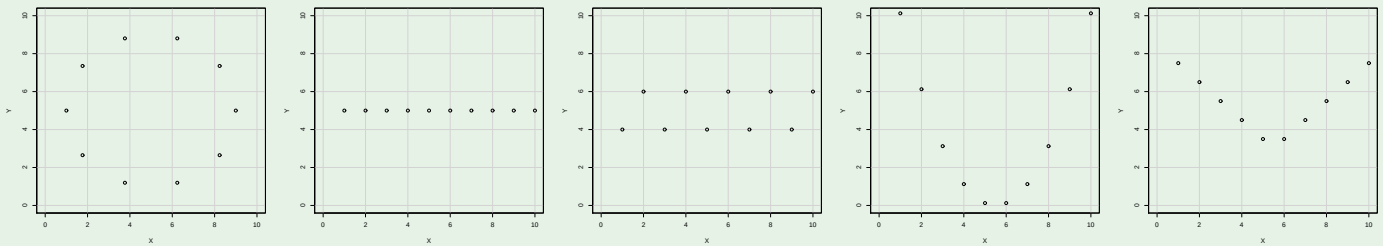
相関係数 $r = 0 \Leftrightarrow x$ と y の間に ‘関係’ がない?

- 相関係数 $r = 0 \Leftrightarrow x$ が増えたら［　　］言えない
- 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-Q8

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)

L04-Q9

Quiz(相関係数)

次のうち, 2 変量データ x_i, y_i の相関係数 r について本当はどれ?

1. x_i をだけ一斉に -2 倍すると, r は -2 倍になる.
2. x_i をだけ一斉に -2 倍すると, r は 2 倍になる.
3. x_i をだけ一斉に -2 倍すると, r は -1 倍になる.
4. x_i をだけ一斉に -2 倍すると, r は $+1$ 倍になる (かわらない).
5. x_i をだけ一斉に -2 倍すると, r は $-1/2$ 倍になる.
6. x_i をだけ一斉に -2 倍すると, r は $1/2$ 倍になる.

にせの因果関係にだまされるな

因果関係=原因と結果
勝利数とシュート回数は正の相関

- 原因:シュートが多い, 結果: 勝利が多い?
- 原因:勝利が多い, 結果:シュートが多い?

(打った) フリーキック回数と被シュート本数は負の相関

- 原因:フリーキックが多い, 結果:被シュートが少ない?
- 原因:被シュートが少ない, 結果:フリーキックが多い?
- 原因:???, 結果:被シュートが少ない, かつ, フリーキックが多い?

- 相関が強くても
- 因果関係があっても相関係数からは原因と結果を区別できない

連絡

- 配布資料は 1-503 向かいの引出, `http://hig3.net` で再配布.
- Quiz の略解は授業終了後に `http://hig3.net` で配布.
- 加減乗除と平方根 (ルート) の使える電卓持ってきてね. 関数電卓でなくてもいいです. 携帯電話の機能・アプリでもかまいません.
- 週のタイムラインで見たように, 非参照 Quiz 予習問題を RaMMoodle に金 17:00 ごろまでに公開. これで来週の Quiz に備えてね.
- 統計検定 申込締切 2015-10-16 金, 受験 2015-11-29 日. 3 級 or 2 級.
- オフィスアワー月 4 木 6(1-502)

manaba 出席カード提出
`https://attend.ryukoku.ac.jp`