箱ひげ図・データの変換・標準得点・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 L03(2017-10-04 Wed)

最終更新: Time-stamp: "2017-10-04 Wed 13:22 JST hig"

今日の目標

- 複数の箱ひげ図, ヒストグラムから分布の性質を記述できる
- データを 1 次関数で標準得点に変換して平均値と分散を比較できる
- 2 変量データの共分散, 相関係数が求められる

http://hig3.net

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 $18\text{cm}.$
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 $Q_2 = x_{(5)}$. よって階級 21–23 に含まれる.
   $x_{(5)} \approx 21 + 2 \times \frac{1.5}{2} = 22.5.$
2. 階級値を答えて, 10
3. $\frac{1}{9}(10 \times 3 + 22 \times 2 + 24 \times 2 + 26 \times 2) = 19.3$

L02-Q3

Quiz 解答:範囲

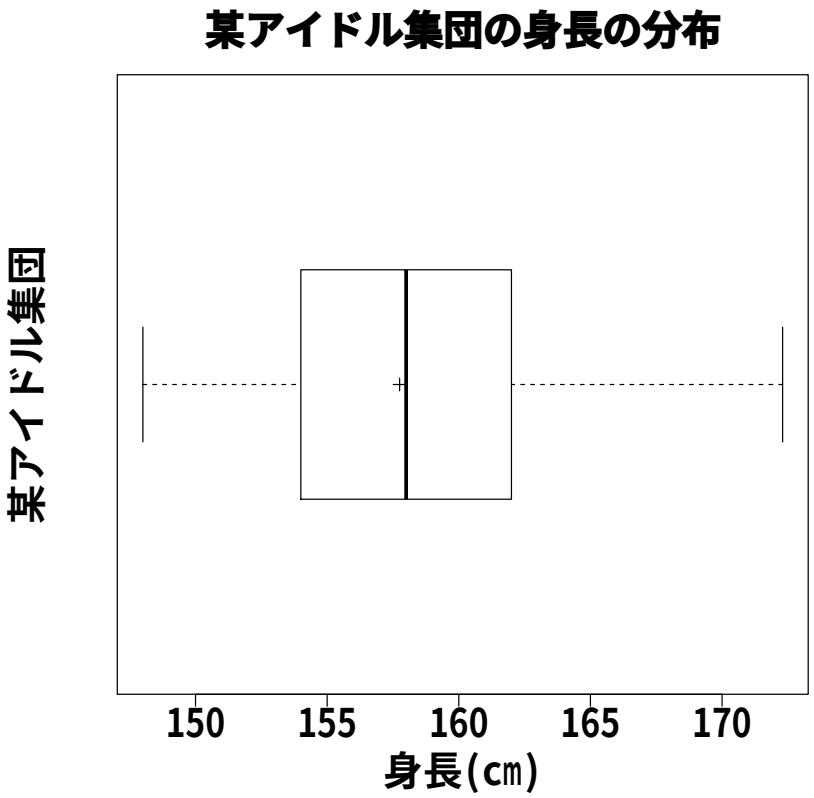
範囲は $Q_4 - Q_0 = 25 - 14 = 11$, 四分位範囲は $Q_3 - Q_1 = 18 - 14.5 = 3.5$, 四分位偏差は $\frac{1}{2}(Q_3 - Q_1) = 1.75$.

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

ここまで来たよ

箱ひげ図 (Box Plot, Box and Whisker diagram)

西川確率統計 p.97



最小最大値　Q_0, Q_4,　四分位点　Q_1, Q_2, Q_3

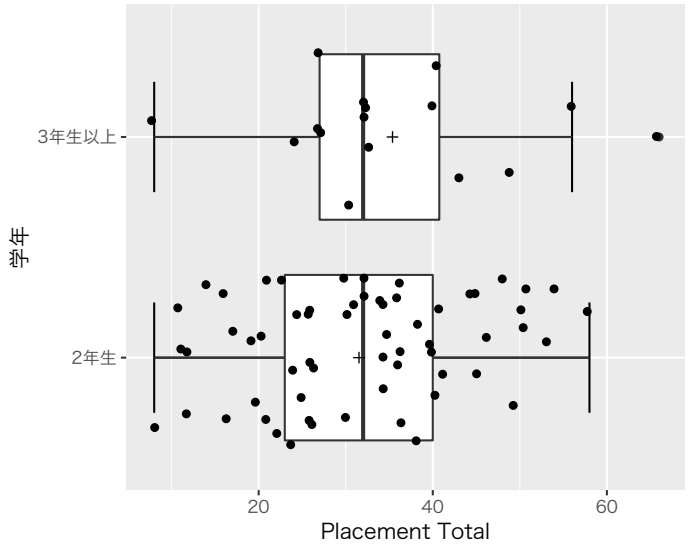
箱ひげ図を描く手順 高校 数学 I

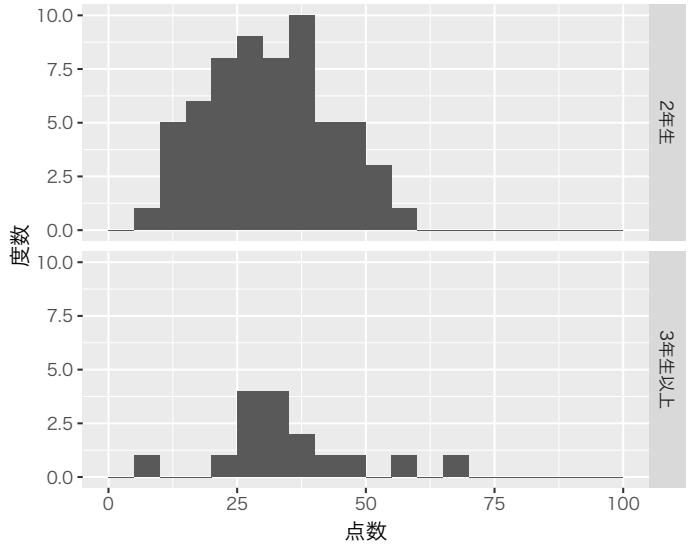
- Q_0, Q_4 Q_1,Q_2,Q_3 と平均値 $\bar{x}$ を求める
- Q_2 に縦線をいれる
- Q_1,Q_3 を左右の端として箱を描く
- Q_0,Q_4 に短い縦線をいれ, 点線のひげで箱とつなぐ
- 平均値に $+$ を 1 個描く

この他に「外れ値を○で描く」こともある.

いまの場合, 横軸:身長 (cm), 縦軸:意味なし

スタートテストの結果





縦軸の意味, ヒストグラムとの使い分け

用語

裾 (すそ,tail) が重い=裾をひいた

右/左に裾が長い=左/右に偏った

ここまで来たよ

分散の意味 I

L03-Q1

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? 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. これだけの情報ではまったくわからない
5. 平均点が 60 点だと再テストがあるだろう

平均値・分散・標準偏差の変換 西川確率統計 §5.1.4

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 西川確率統計定理 5.2(p.101)

$y = ax + b$ 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-Q2

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

身長と靴のサイズじゃ標準偏差の意味が違う!

西川確率統計 §5.1.5

Berryz 工房内で, 身長の標準偏差は 20cm くらいだけど, 靴のサイズの標準偏差は 3cm くらい.

標準偏差が大きい = いろんな体格の人がいる

みたいに思いたいけど, 身長と靴のサイズじゃ標準偏差の意味が違う.

変動係数 (coefficient of variation)

$$(\text{データ } x \text{ 全体の}) \text{ 変動係数} = \frac{S_x}{\overline{x}} \times 100$$

- これは無次元の数. すなわち単位がない量.

  .
- $\frac{\text{分散}}{\text{平均値}}$ だと無次元の数にはならない.

標準得点

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 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.

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, 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)

L03-Q4

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.

ここまで来たよ

2. 略解:データの代表値・散らばりの尺度

3. 箱ひげ図・データの変換・標準得点・2 変量データ
   - 箱ひげ図
   - 分散の意味と平均値・分散・標準偏差の変換
   - 変動係数・標準得点・偏差値

4. 2 変量データ
   - 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).

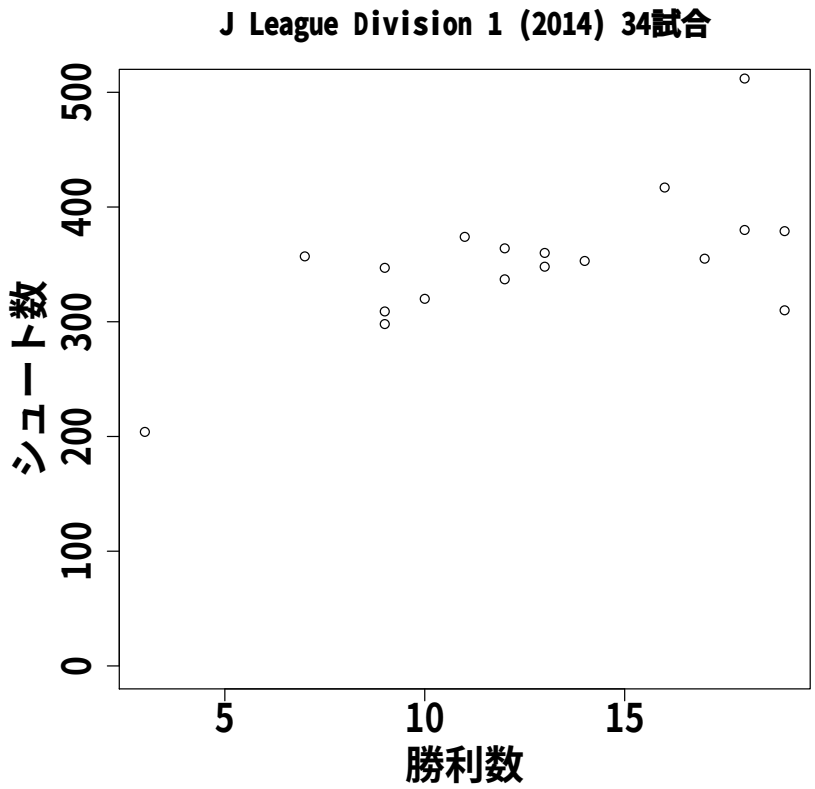
(チーム名)	x	y	z
ベガルタ仙台	9	347	50
鹿島アントラーズ	18	512	39
⋮	⋮	⋮	⋮
計	...	...	...
平均値	...	...	...

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=相関図

西川確率統計 §5.2.2



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.

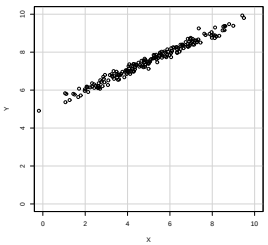
$\downarrow y$　$\backslash x$ の階級 $\rightarrow$	0 以上 5 未満	10 未満	15 未満	20 未満	計
200 以上 250 未満	1				1
250 以上 300 未満		1			1
300 以上 350 未満		2	3	1	6
350 以上 400 未満		1	4	3	8
400 以上 450 未満				1	1
450 以上 500 未満				0	0
500 以上 550 未満				1	1
計	1	4	7	6	18

周辺分布とは

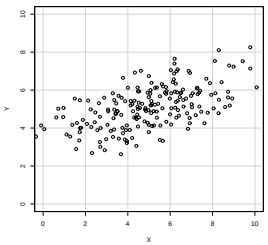
ここまで来たよ

正の相関・負の相関・無相関

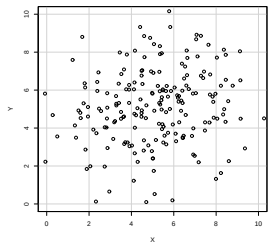
西川確率統計 §5.2.3



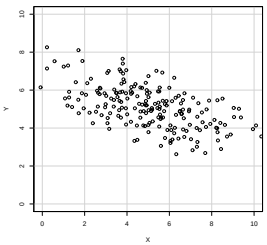
強い正の相関
$r = 0.99$



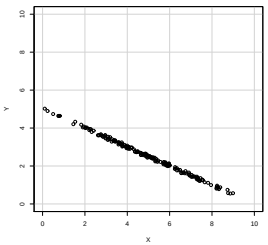
弱い正の相関
$r = 0.55$



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 高校 数学 I 発展 西川確率統計 §5.2.3

相関の強さを数で表したい

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

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

$\overline{y}, S_y^2$ も同様.

共分散 (covariance)

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

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.

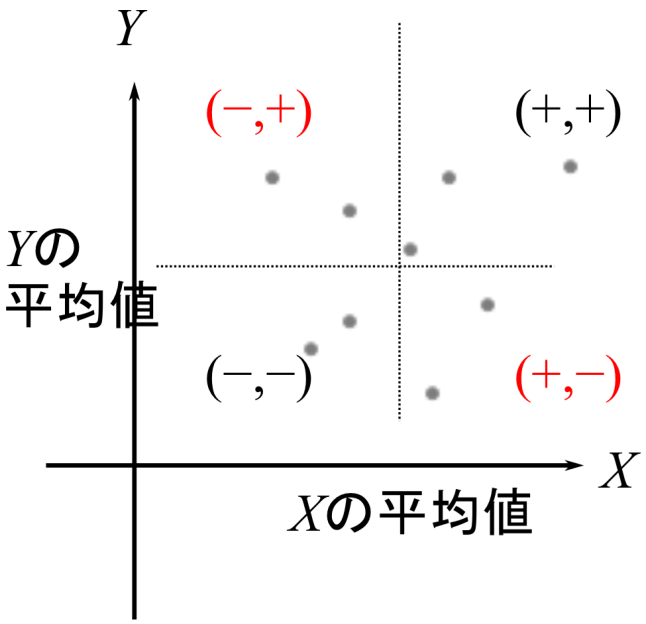
L03-Q5

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 西川確率統計 p.110



$(+,-) = (x_i - \overline{x}$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　西川確率統計 p.111

共分散は

- x, y の 1 次関数による変換で変わる　西川確率統計定理 5.4(p.112)
- 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
- $-1 \leq r \leq +1$ 西川確率統計定理 5.5(p.114)
- $r = 0 \Leftrightarrow$ '無相関' しかし…(待て次回)
- $r = \pm 1 \Leftrightarrow$ 散布図の点が傾き正/負の一直線上 $\Leftrightarrow$ y は x の 1 次関数.
  西川確率統計定理 5.7(p.115)
- r は x, y の 1 次関数による変換のもとで不変 西川確率統計定理 5.6(p.114)

連絡

- 次回は 1-609 実習室. 動画見ます. イヤフォン持ってきて.
- Excel 使います. 慣れてない人は Excel 入門コースで第 4 章 2 までやっておいて. `https://moodle.media.ryukoku.ac.jp`
- 配布資料は 1-503 向かいの引出や `http://hig3.net` で再配布.
- 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.
- Learn Math Moodle の予習復習問題で来週の trial に備えてね.
- 樋口オフィスアワー月 3.5(1-539) 金 4(1-502), Math ラウンジ月-木昼 (1-614)
- 来週は教科書 西川確率統計 5.2.4, 5.2.5, 5.2.6 読んできて

統計検定のディスカウント受験受付中 (– 2017-10-09 月) 樋口まで. 3 級合格者はプチテストの点数の一部として使用可.