2 変量データの共分散・相関係数・回帰分析

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(2017-10-11 Wed)

最終更新: Time-stamp: "2017-10-10 Tue 23:02 JST hig"

今日の目標

- 2 変数の量的データから, 手で共分散と相関係数と回帰直線が求められる
- 1 変数の量的データから, Excel で代表値・分散が求められる
- 2 変数の量的データから, Excel で共分散と相関

http://hig3.net

L03-Q1

L03-Q2

Quiz 解答:平均値・分散・標準偏差の換算

1.6m, 0.0025m^2, 0.05m.

L03-Q3 Quiz 解答:標準得点と偏差値

平均値 $\overline{x} = 90$, 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

1. 箱ひげ図・データの変換・標準得点
2. 2 変量データの共分散・相関係数・回帰分析
   - 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
   - 回帰分析
   - Excel で統計

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).

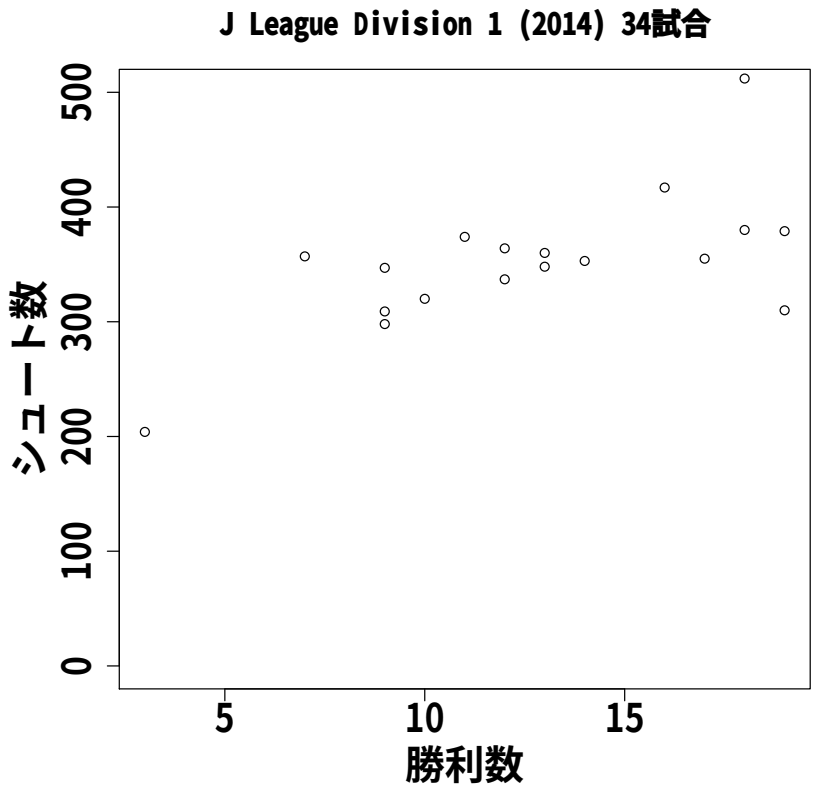
(チーム名)	x	y	z
ベガルタ仙台	9	347	50
鹿島アントラーズ	18	512	39
⋮	⋮	⋮	⋮
計	...	...	...
平均値	...	...	...

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

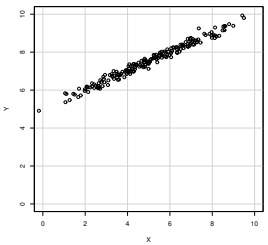
周辺分布とは

ここまで来たよ

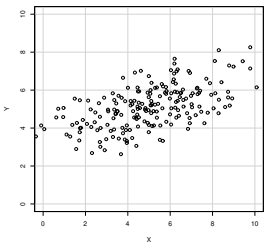
1. 箱ひげ図・データの変換・標準得点
2. 2 変量データの共分散・相関係数・回帰分析
   - 2 変量データとクロス集計表・散布図
   - 2 変量データの相関
   - 回帰分析
   - Excel で統計

正の相関・負の相関・無相関

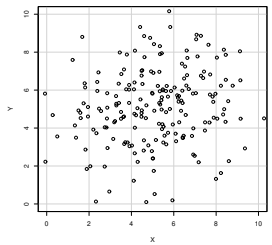
西川確率統計 §5.2.3



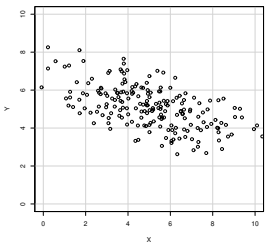
強い正の相関
$r = 0.99$



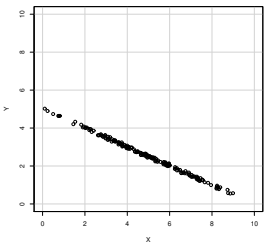
弱い正の相関
$r = 0.55$



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 r_{xy} とも書く. 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

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

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

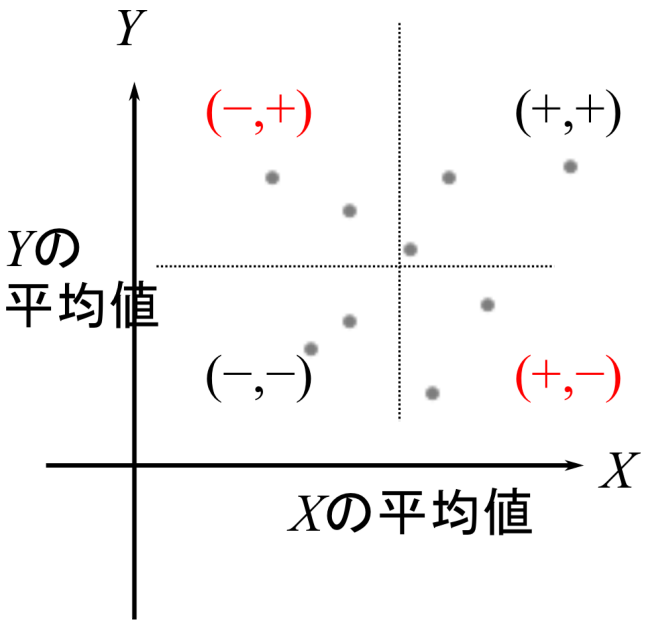
$\overline{y}, S_y^2$ も同様.

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

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)

L04-Q1

Quiz(共分散と相関係数 (単位付き))

次の $(x\mathrm{g}, y\mathrm{cm})$ のデータがある

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 $= \sqrt{\frac{122}{5}} = 4.94(\mathrm{cm})$ は使っちゃっていい.

$x(\mathrm{g})$	$y(\mathrm{cm})$
1	5
3	15
4	14
5	11
7	20

ここまで来たよ

1. 箱ひげ図・データの変換・標準得点
2. 2 変量データの共分散・相関係数・回帰分析
   - 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
   - 回帰分析
   - Excel で統計

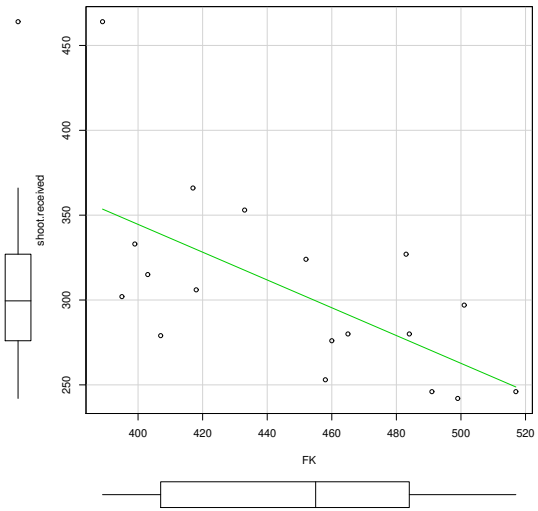
回帰分析 西川確率統計 §5.2.4

回帰 (regression), 直線回帰=単回帰分析=1 変数回帰分析　物理実験

2 変量データ (x, y) が
相関係数 $r = \pm 1$ に近い $\Leftrightarrow$ 散布図上のデータ点 (x, y) がほぼ直線に載っている

その直線 (　　) の式 $y = ax + b$ を知りたい!

つまり　　a, 定数項 b を決めたい.



y: 目的変数 (従属変数)
x: 説明変数 (独立変数)
何でそんなことしたいの?

- 法則を見つけたい
- x から y を予測したい

回帰直線の決め方

1. 定規をあてて ‘真ん中’ を通るように
2. 最小 2 乗法で.

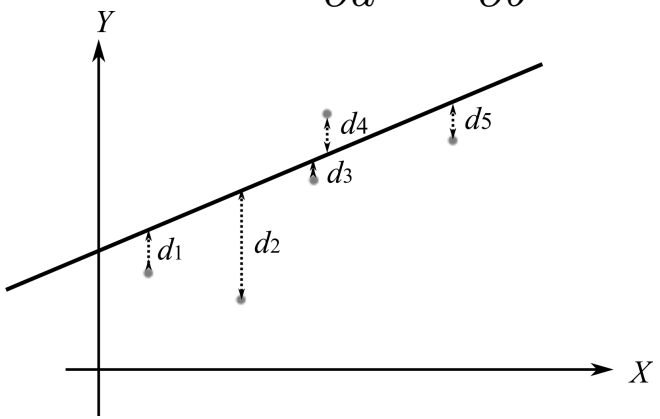
最小 2 乗法

直線からのずれの 2 乗 d^2 の合計

$$L(a,b) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

の最小条件 $\frac{\partial L}{\partial a} = \frac{\partial L}{\partial b} = 0$ で a, b を決める.

微積分 I



物理実験

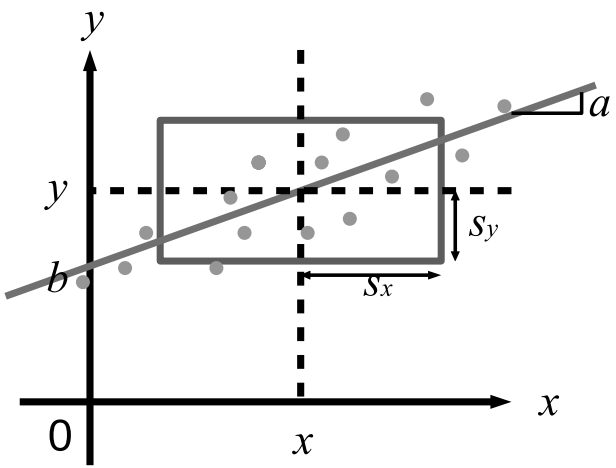
直線回帰の公式

回帰直線 西川確率統計定理 5.8, 式 (5.11)

x_i, y_i $(i = 1, \ldots, n)$ の平均値を $\overline{x}, \overline{y}$, 標準偏差を S_x, S_y, 相関係数を r とする. このとき回帰直線は,

$$y = \frac{r \times S_y}{S_x} \times (x - \overline{x}) + \overline{y} = ax + b.$$

傾きは $a = \frac{r \times S_y}{S_x} = \frac{C_{xy}}{S_x^2}$, 切片は $b =$ (点 $(\overline{x}, \overline{y})$ を通るような値)



a: 回帰係数 (x を 1 だけ変えたときの y の変化量)
r^2: 決定係数 (あてはまりのよさ) 西川確率統計 §5.2.4

誤差

回帰直線の傾きのおぼえ方 I

広がり方

散布図上のデータ点の分布は, 横 $2S_x$, 縦 $2S_y$ $\rightarrow$ 傾き $\frac{S_y}{S_x}$ くらい?
しか〜し, 傾きには正負があるし, 相関がなかったら傾きを 0 にしたいので, 相関係数 r をかけ算しておく.

単位チェック

(x, y) の単位が (m,kg) だとする.
r は無次元. 単位無し.
左辺 y (kg).
右辺 $r \times \frac{S_y(\mathsf{kg})}{S_x(\mathsf{m})} \times x(\mathsf{m}) + b(\mathsf{kg})$
で, S_x/S_y かけると単位があう.

L04-Q2

Quiz(回帰係数と回帰直線)

ある 2 変量データ (x, y) について次のことがわかっている.

x の平均値 $\overline{x}$	9
y の平均値 $\overline{y}$	-4
x の分散 s_x^2	49
y の分散 s_y^2	36
x, y の共分散 s_{xy}	-25
(x, y) のデータの個数 n	16

このとき, 回帰直線の式を, x, y の式で書こう. 整理しなくてよい.

ここまで来たよ

1. 箱ひげ図・データの変換・標準得点

2. 2 変量データの共分散・相関係数・回帰分析
   - 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
   - 回帰分析
   - Excel で統計

準備

統計ソフトウェア実習室にインストールされているのは

R 無料. オープンソース. 解説書が多い.
SPSS 伝統ある高級品.
Excel 機能は限られ怪しいところもあるが, 普及率高い. 龍大では Office365 で無料.

今日は Excel を使ってみます.
スタートボタン >Excel 2016
統計分析のための準備
ファイル > オプション > アドイン > Excel のアドイン > 設定 > 分析ツール にチェックを入れて OK する.

表計算ソフトウェア (Excel) による主な分析 高校 数学 I

どこかの段階でデータ範囲を指定, または関数の引数にデータ範囲を指定.

	メニューベース	関数ベース
平均値, 分散, 標準偏差	データ > 分析 > データ分析 > 基本統計量 > 統計情報	平均値 average, 分散 var.p, 標準偏差 stdev.p, 最頻値 mode
四分位数	データ > 分析 > データ分析 > 順位と百分位数	中央値 median, 四分位数 quartile
度数分布表, ヒストグラム	データ > 分析 > データ分析 > ヒストグラム > 入力範囲とデータ区間	frequency + グラフ
散布図	挿入 > グラフ > 散布図	
共分散, 相関係数	データ > 分析 > データ分析 > 共分散, 相関	covar=covariance.p, correl
回帰分析	データ > 分析 > データ分析 > 回帰分析	linest
クロス集計表	挿入 > テーブル > ピボットテーブル	

行=横のセル

の並び, 列=縦のセルの並び

メニューベースのデータ分析¿基本統計量の分散は, さらに $\frac{n-1}{n}$ 倍しないと, 「データの分散」var.p にならない.

メニューベースの分析をするときの注意

- Excel は, 1 種類のデータは列方向 (縦方向) にならんでいるとデフォルトでは想定する. 分析の種類によっては, 列方向, 行方向のどちらに並んでいるかを指定できるものもある.
- 2 変量 (n 変量) の統計量である, 共分散 S_{xy} や相関係数 r_{xy} の出力は

$$\begin{matrix} S_{xx} & S_{yx} \\ S_{xy} & S_{yy} \end{matrix}, \qquad \begin{matrix} r_{xx} & r_{yx} \\ r_{xy} & r_{yy} \end{matrix}$$

のように行列状ににになっている. S_{yy} や r_{yy} は, $y = x$ であるときの S_{xy}, r. よく考えると, $S_{yy} = S_y^2, r_{yy} = 1$ であることに気づく. $n \geq 3$ のときは $n \times n$ 行列になる.

- 回帰分析の出力では
  - 重相関 R $=$ 相関係数 r
  - 重決定 R2 $=$ 決定係数 r^2
  - 切片の係数 $=$ 回帰直線の切片 b
  - X 値 1 の係数 $=$ 回帰係数 a
  - $n \geq 3$ の重回帰 $(x_1, x_2, \ldots, x_{n-1}, y)$ というものがあり, そのときは X 値 2,$\cdots$ などとなっていく.

連絡

- 2017-10-18 水 は全学休講
- 2017-10-25 水 は, 今回の内容に相当する trial はありません. そのかわり予習復習問題が Excel が必要なヘビーなものになる予定. 締切 2017-10-25 水 23:59
- 2017-10-25 水は, 事前に各自で動画で学習, 授業中に演習, 授業の最後に (その日の分の) trial となる予定.
- 配布資料は 1-503 向かいの引出, `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.
- 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)
- 次回は 西川確率統計 §1.4,§2.1,§2.2,§2.3 から内容の一部分を選択して進みます.