



統計学入門 第6回

早稲田大学政治経済学部
西郷 浩



本日の目標

- 2次元データの分析
 - 散布図と相関
 - 相関を測る尺度
 - 分割表
 - PC実習

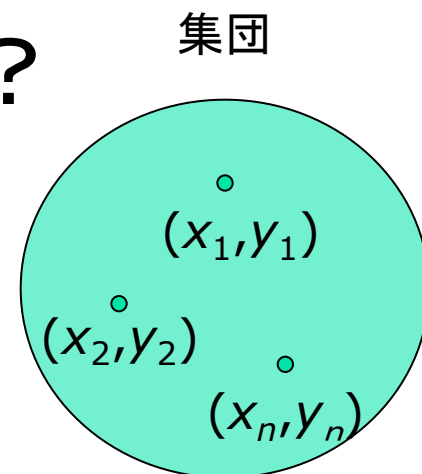
関係の分析(1)

- 2次元データ

- $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

- どのように分析すべき？

- x のみ(y のみ) → 可能
- (x, y) を同時に扱う
 - 関係の分析



関係の分析(2)

■ 2次元分布の表示

- 散布図：
 - データを x - y 平面上に表示
- 分割表：
 - 多次元度数分布表

表1: 2次元データの要約方法

x / y	数量	属性
数量	散布図 分割表	分割表
属性	分割表	分割表

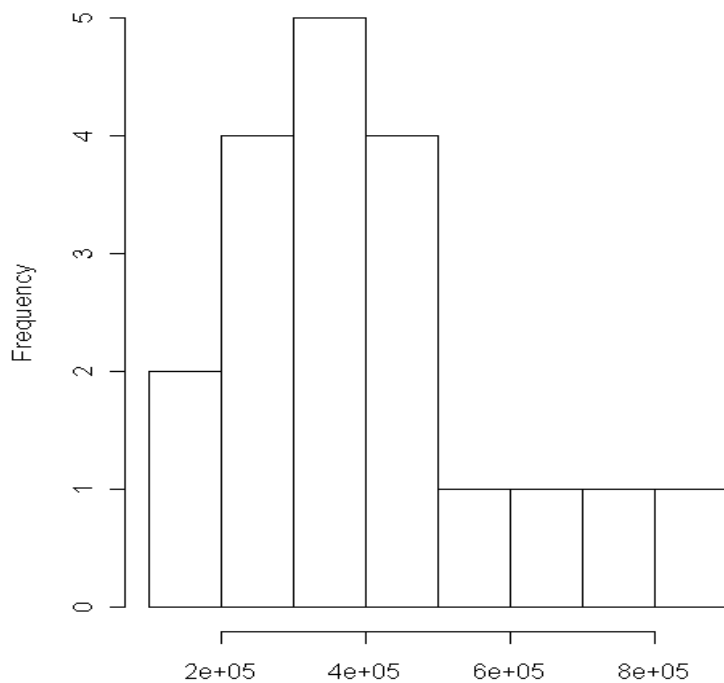


2次元データの例(1)

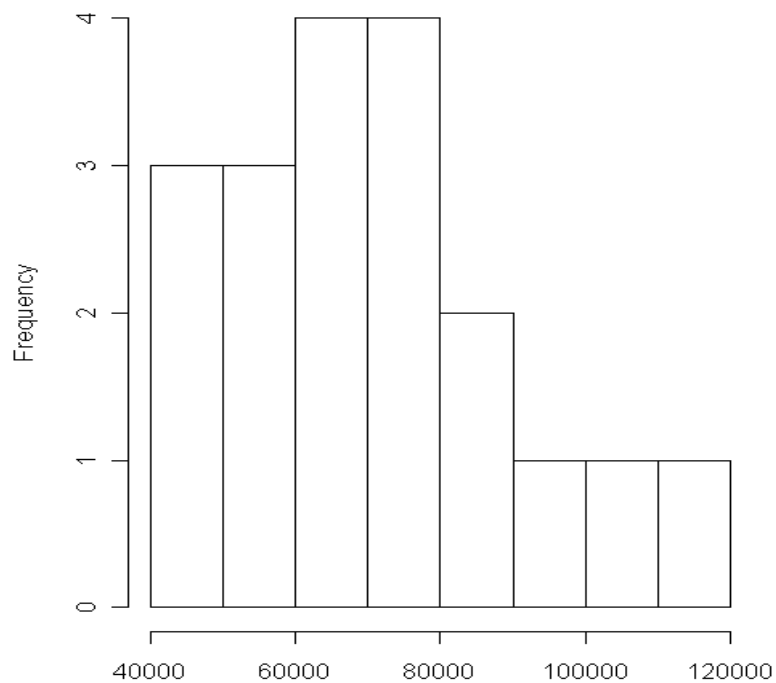
- 総務省統計局
「平成21年全国消費実態調査」
 - 表1 年間収入階級別一世帯当たり1か月の収入と支出(2人以上世帯のうち勤労者世帯)
 - 可処分所得(x), 食料(y)
- 1次元データとしての分析
 - ヒストグラム

2次元データの例(2)

図1: 可処分所得と食料への支出のヒストグラム



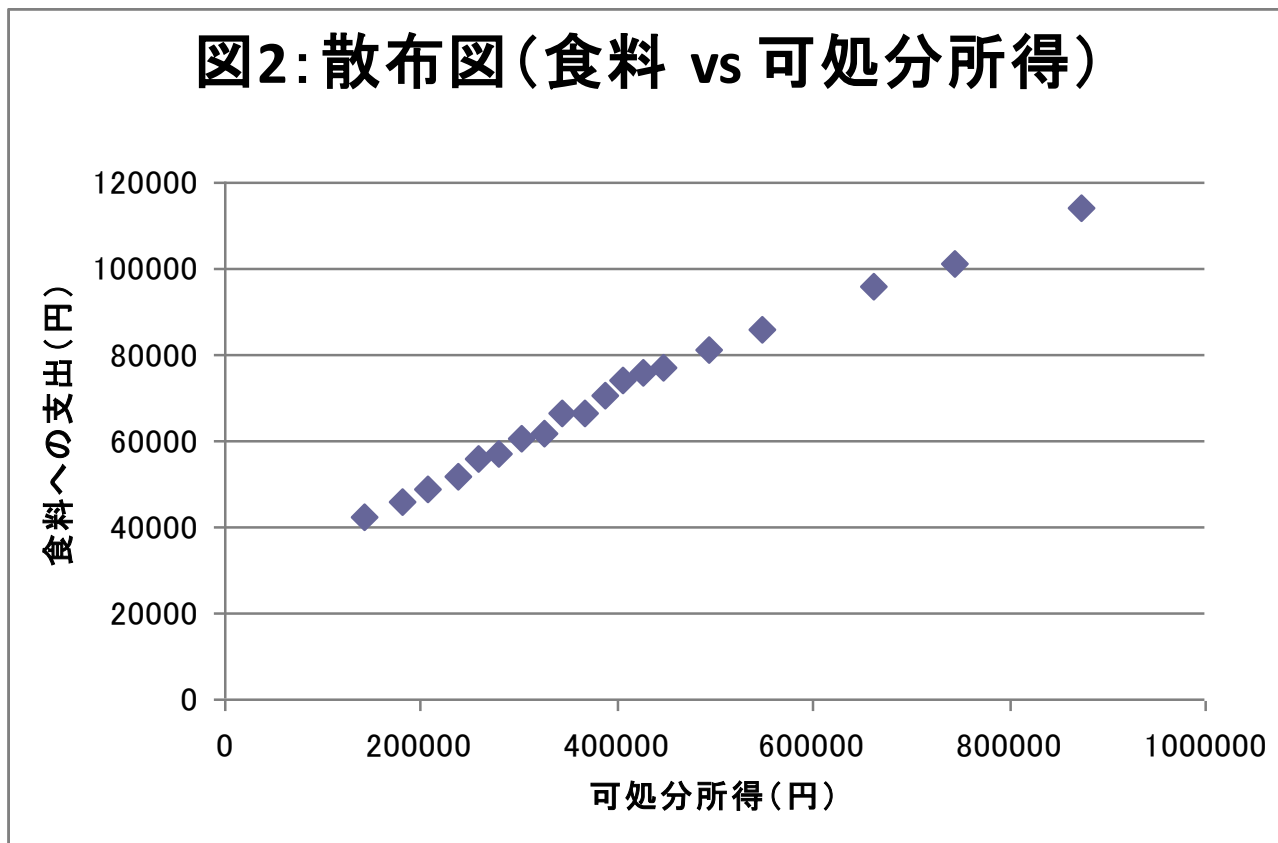
(a)income



(b)food

資料: 総務省統計局「平成21年全国消費実態調査」表1

散布図(1)



資料: 総務省統計局「平成21年全国消費実態調査」表1



散布図(2)

- 散布図から読み取れること
 - 右上がりの傾向
 - 可処分所得 \uparrow (\downarrow) \Leftrightarrow 食料 \uparrow (\downarrow)
 - 直線関係の強弱
 - ほぼ一直線。しかし、厳密には直線でない。



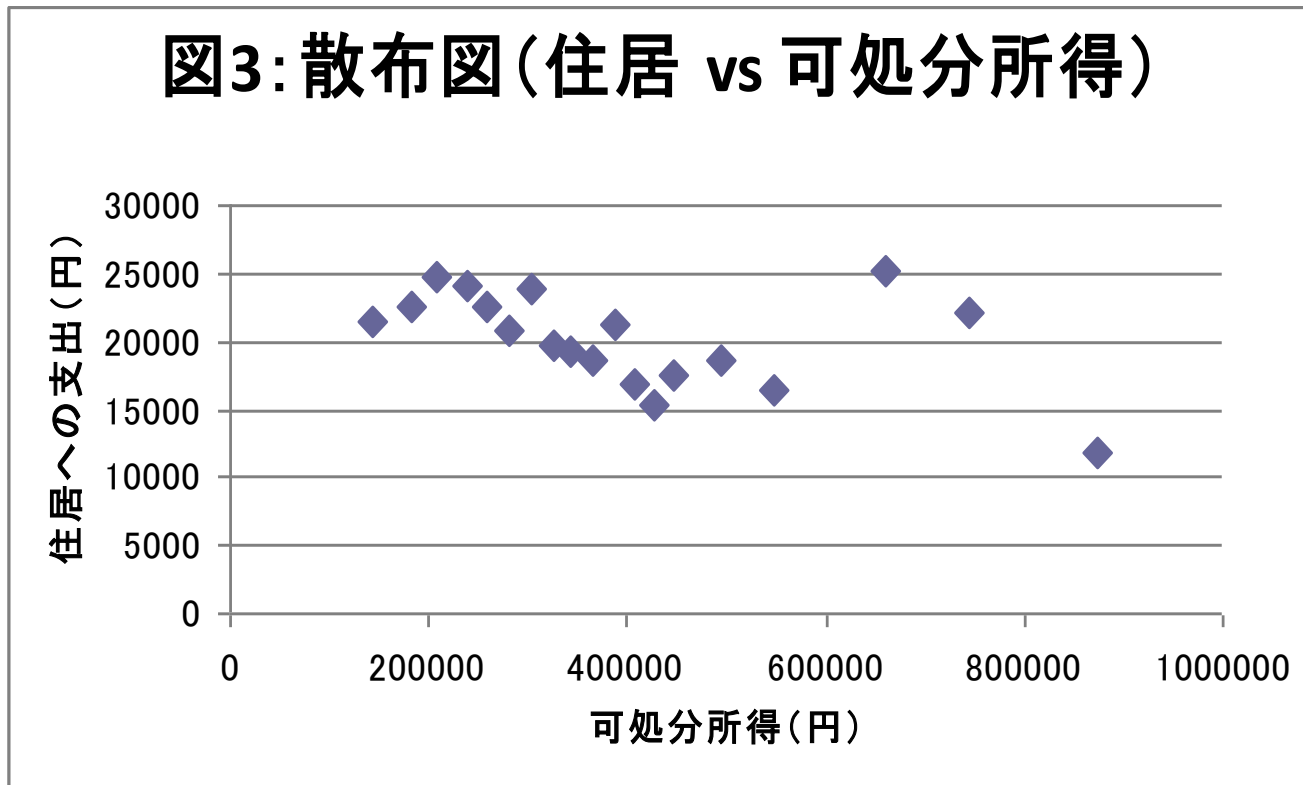
相関(1)

■ 相関

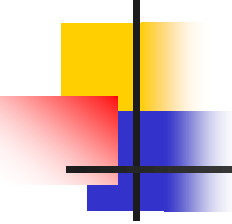
- ふたつの変数 x, y の直線関係の強さ
 - 強い正の相関: 右上がりの直線関係
 - 弱い正の相関: 右上がりの傾向
 - 無相関: はっきりした傾向なし
 - 弱い負の相関: 右下がりの傾向
 - 強い負の相関: 右下がりの直線関係

相関(2)

図3: 散布図(住居 vs 可処分所得)

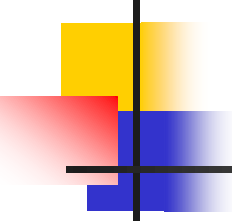


資料: 総務省統計局「平成21年全国消費実態調査」表1



相関を測定するための尺度(1)

- 散布図による相関の把握
 - 有効 but 主観的
- 数値化の必要性
 - 共分散: S_{xy}
 - 相関係数: r_{xy}



相関を測定するための尺度(2)

共分散

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

共分散の符号と相関の符号

$S_{xy} > 0 \Leftrightarrow$ 相関が正 \Leftrightarrow 散布図が右上がり

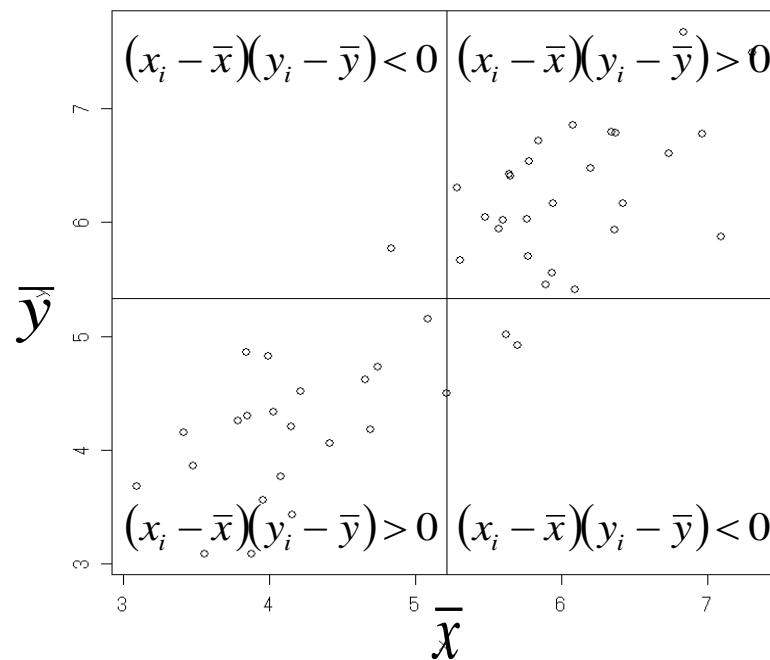
$S_{xy} \approx 0 \Leftrightarrow$ 相関ない \Leftrightarrow 明確な傾向なし

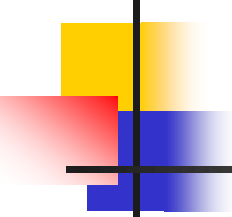
$S_{xy} < 0 \Leftrightarrow$ 相関が負 \Leftrightarrow 散布図が右下がり

相関を測定するための尺度(3)

- 平均からの偏差の積の符号
 - 散布図右上がり
 - プラスが多い
 - $S_{xy} > 0$ となる。
(右下がりのときはマイナスが多くなる。)

図4: 共分散の符号





相関を測定するための尺度(4)

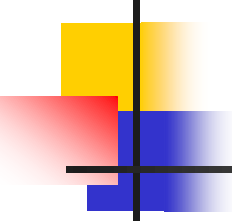
- 可処分所得 x と食料 y との共分散
 - $S_{xy} = 3,527,329,662$
 - プラスになるので、散布図に見られる右上がりの傾向と合致している。
 - But 関係の強弱をあらわしているだろうか？
 - たとえば、測定単位を千円単位に変更したら？
 - 測定単位を変更しても、「 x と y との関係自体に変わりはない」と考えるのが自然である。
- 共分散を「標準化」する必要性
 - 変数の測定単位と無関係な無名数が好ましい。



相関を測定するための尺度(5)

相関係数

$$\begin{aligned} r_{xy} &= \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$



相関を測定するための尺度(6)

- 相関係数の性質

- $-1 \leq r_{xy} \leq 1$

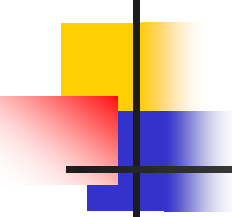
- 強い正の相関 $\Leftrightarrow r_{xy} \doteq 1$

- 正の相関 $\Leftrightarrow 0 < r_{xy} < 1$

- 無相関 $\Leftrightarrow r_{xy} \doteq 0$

- 負の相関 $\Leftrightarrow -1 < r_{xy} < 0$

- 強い負の相関 $\Leftrightarrow r_{xy} \doteq -1$



相関を測定するための尺度(7)

■ 相関係数の値

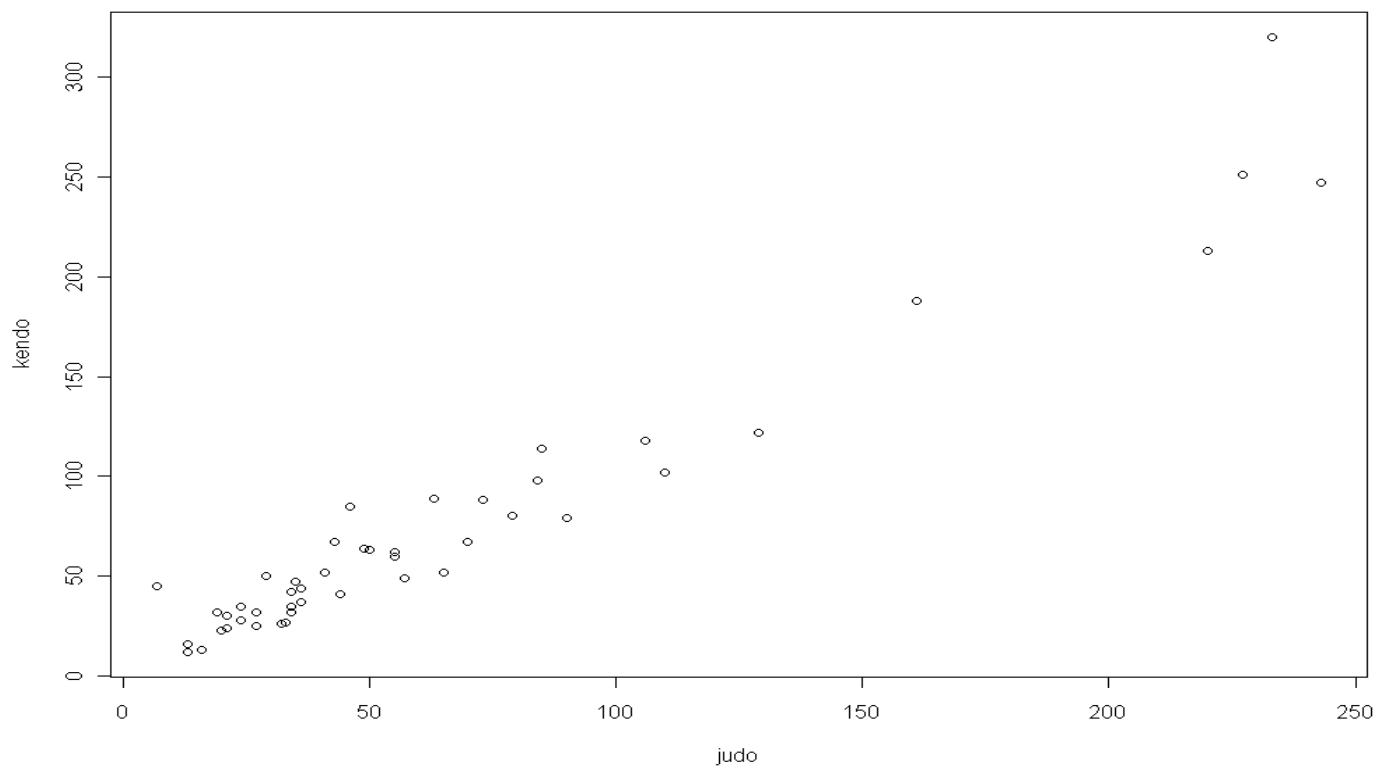
- 可処分所得と食料(図2): $r_{xy} = 0.99$
- 可処分所得と住居(図3): $r_{xy} = -0.48$

■ 注意点

- 直線関係の強弱を示すのみ。
- 「強い相関関係→因果関係」とは限らず。
 - 因果関係を主張するためには、理論的な背景が必要になる。

相関を測定するための尺度(8)

図5: 都道府県別剣道場数と柔道場数



資料: 総務省統計研修所編(2011)『第61回日本統計年鑑』表23-15

分割表 (1)

同時分布(結合分布)

表2: 可処分所得と住居への支出の分割表

表頭

可処分所得を所与としたときの
住居に関する条件つき分布

(住居についての)周辺分布

表側

住居への支出	可処分所得			合計
	0万-29.9万	30万-59.9万	60万-89.9万	
10千-14.9千	0	0	1	1
15千-19.9千	0	8	0	8
20千-24.9千	6	2	1	9
25千-29.9千	0	0	1	1
合計	6	10	3	19

資料: 総務省統計局「平成21年全国消費実態調査」表1



分割表 (2)

- 2つの変数の関係
 - 同時分布(結合分布)
 - 条件つき分布
 - 所与(条件)とした変数の値を変化させると、2つの変数の関係がわかる。
- 相対度数による表示
 - 行和(列和)に対する相対度数。

分割表 (3)

表3: 2011年度入学者 学部(X)と性別(Y)の同時分布

(a) 度数を表示した分割表

$x \backslash y$	男	女	合計
政経	708	312	1020
文文	745	941	1686
合計	1453	1253	2706

(b) 行和に対する相対度数

$x \backslash y$	男	女	合計
政経	0.69	0.31	1.00
文文	0.44	0.56	1.00
合計	0.54	0.46	1.00

資料: 早稲田大学教務部「統計で見る早稲田大学」2011年度版



分割表 (4)

- 質的変数どうしの分割表
 - 変数の順序に大小・高低の意味がない場合、「相関」の定義を工夫する必要がある(一般の場合は複雑になる)。
- 2×2 の分割表のための関連係数
 - 相関係数に対応するもの。
 - ただし、変数の順序に大小・高低の意味がないときには、符号は無意味。



分割表 (5)

表4: 2×2 の分割表

$x \backslash y$	G	H	行和
E	a	b	$a+b$
F	c	d	$c+d$
列和	$a+c$	$b+d$	n

関連係数

$$R = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

分割表 (6)

表5: 人工的な例 ($R=0$ となる)

$x \backslash y$	G	H	行和
E	4	6	10
F	8	12	20
列和	12	18	30

表6: 人工的な例 ($R=1$ となる)

$x \backslash y$	G	H	行和
E	10	0	10
F	0	20	20
列和	10	20	30

学部・性別データ

$$R=0.24$$



PC実習

- 散布図の作成
- 共分散・相関係数の計算
- 分割表の作成