

統計の分析と利用

1. データとその扱い

堀田 敬介

1-1. 一次元のデータ

- 1変数の図化: 度数分布とヒストグラム, 幹葉プロット, 箱ひげ図
- 1変数の数表現: 代表値と散らばり, データの標準化

2-2. 二次元のデータ

- 2変数の関係1: 散布図, 共分散・相関係数
- 2変数の関係2: クロス集計, クラメアの連関係数
- 2変数の関係3: 点グラフ, 相関比

2010/10/8, Fri. ~

2次元のデータ

□ 2変数 x, y の関係

因果関係が分かるわけではない

□ 相関 correlation

- x と y との間に区別をつけず対等に見る見方・方法, 単なる関係
- 例: 数学の成績と英語の成績

	A	B	C	D	E	F	G	H	I
数学 x	65	75	84	72	69	70	72	68	78
英語 y	59	68	75	72	69	65	60	68	74

□ 回帰 regression

- x から y を見る見方・方法
- ある一方が他方を左右する場合
- 例: 年齢と血圧, 所得と消費, 人口と商業, 気候と住環境

血圧

年齢

2次元のデータ

2変数 x, y の関係

		尺度									
例1		A	B	C	D	E	F	G	H	I	J
性別 x		男	男	女	男	男	男	女	女	男	女
嗜好 y		紅茶	緑茶	珈琲	珈琲	緑茶	珈琲	紅茶	珈琲	珈琲	紅茶
		クロス集計					連関係数				
例2		A	B	C	D	E	F	G	H	I	J
飲量 x		15	32	16	30	51	11	14	24	15	19
嗜好 y		紅茶	緑茶	珈琲	珈琲	緑茶	珈琲	紅茶	珈琲	珈琲	紅茶
		点グラフ					相関比				
例3		A	B	C	D	E	F	G	H	I	J
身長 x		176	170	163	173	170	171	165	170	176	156
体重 y		61	73	54	65	67	62	51	57	77	43
		散布図					相関係数				

2変数の関係

2変数の関係1: x (質的) $\times$ y (質的)

	A	B	C	D	E	F	G	H	I	J	
性別 x	男	男	女	男	男	男	女	女	男	女	質的
嗜好 y	紅茶	綠茶	珈琲	珈琲	綠茶	珈琲	紅茶	珈琲	珈琲	紅茶	質的
└─┬─┐ クロス集計											
	紅茶	綠茶	珈琲	計							
男	1	2	3	6	} 周辺度数						
女	2	0	2	4							
計	3	2	5	10	← 総度数						
} 周辺度数											

2変数の関係

□ 2変数の関係1: x (質的) $\times$ y (質的)

	紅茶	緑茶	珈琲	計
男	1	2	3	6
女	2	0	2	4
計	3	2	5	10

連関係数

クロス集計
から
理論度数
を
求める

	紅茶	緑茶	珈琲	計
男	1.8	1.2	3.0	6
女	1.2	0.8	2.0	4
計	3	2	5	10

$$1.8 = \frac{3 \cdot 6}{10}$$

$$2.0 = \frac{5 \cdot 4}{10}$$

□ クラメル の 連関係数 *Cramer's coefficient of association*

$$V = \sqrt{\frac{\chi^2}{n \cdot m}}$$

$$(0 \leq V \leq 1)$$

$$\chi^2 = \frac{(1-1.8)^2}{1.8} + \frac{(2-1.2)^2}{1.2} + \dots + \frac{(0-0.8)^2}{0.8} + \frac{(2-2.0)^2}{2.0}$$

$$n = 10$$

$$m = \min\{2-1, 3-1\}$$

ピアソンの
 χ^2 統計量

(行数-1)と(列数-1)
の小さい方

2変数の関係

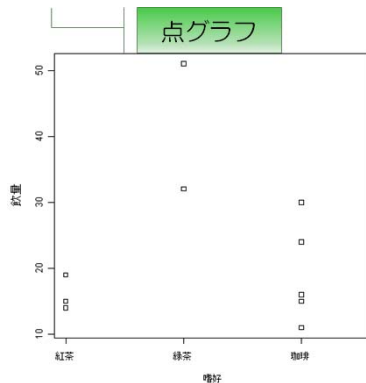
□ 2変数の関係2: x (量的) $\times$ y (質的)

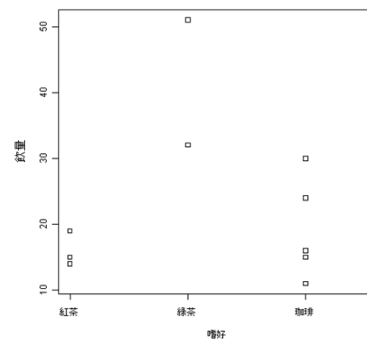
	A	B	C	D	E	F	G	H	I	J
飲量 x	15	32	16	30	51	11	14	24	15	19
嗜好 y	紅茶	緑茶	珈琲	珈琲	緑茶	珈琲	紅茶	珈琲	珈琲	紅茶

量的

質的

点グラフ





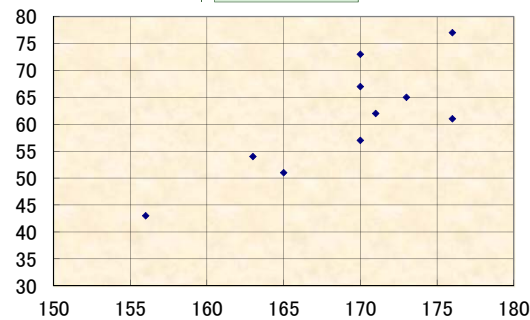
	紅茶	緑茶	珈琲	
	14	32	11	
	15	51	15	
	19		16	
			24	
			30	
個数	3	2	5	全平均
平均	16.0	41.5	19.2	22.7
偏差平方和	44.9	353.4	12.3	$902.8 = S_T$
偏差平方和	4.0	90.3	67.2	級間変動
	1.0	90.3	17.6	
	9.0		10.2	
			23.0	
			116.6	
			合計	
計	14.0	180.5	234.8	$429.3 = S_B$

2変数の関係

□ 2変数の関係3: x (量的) $\times$ y (量的)

	A	B	C	D	E	F	G	H	I	J	
身長 x	176	170	163	173	170	171	165	170	176	156	量的
体重 y	61	73	54	65	67	62	51	57	77	43	量的

散布図



2変数の関係

□ 2変数の関係3: x (量的) $\times$ y (量的)

	A	B	C	D	E	F	G	H	I	J	平均
身長 x	176	170	163	173	170	171	165	170	176	156	169
体重 y	61	73	54	65	67	62	51	57	77	43	61

相関係数

□ ピアソンの積率相関係数 *Pearson's product-moment correlation coefficient*

$$r_{xy} = \frac{\text{cov}_{xy}}{S_x \cdot S_y} \quad (-1 \leq r_{xy} \leq 1)$$
$$\left\{ \begin{array}{l} \text{cov}_{xy} = \frac{(176-169)(61-61) + \dots + (156-169)(43-61)}{10} \quad (x, y \text{の共分散}) \\ S_x = \sqrt{\frac{(176-169)^2 + \dots + (156-169)^2}{10}} \quad (x \text{の標準偏差}) \\ S_y = \sqrt{\frac{(61-61)^2 + \dots + (43-61)^2}{10}} \quad (y \text{の標準偏差}) \end{array} \right.$$

2変数の相関 x (量的) $\times$ y (量的) についての補足
ピアソンの積率相関係数に関する補足と注意点

2変数の相関に関する補足

2変数の相関

□ 共分散 *covariance*

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

(2次元データ $\{x_1, \dots, x_n\}, \{y_1, \dots, y_n\}$ について)

ある i 番目のデータについて、 x_i と平均 $\bar{x}$ との差と、 y_i と平均 $\bar{y}$ との差が共に大きいとき、共分散の値は大きくなり、そうではないとき共分散の値は小さくなる。すなわち、2種類のデータの関係の強さを表している。

□例：文教太郎君と湘南花子さんの昼食に掛けた費用

	月	火	水	木	金
太郎	¥400	¥300	¥100	¥200	¥200
花子	¥100	¥200	¥300	¥400	¥200

太郎君がリッチな食事をとるとき、花子さんは貧乏な食事で我慢してるの？

2変数の相関

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

□ 共分散 *covariance*

□ 例: 文教太郎君と湘南花子さんの昼食に掛けた費用

	月	火	水	木	金
太郎	¥400	¥300	¥100	¥200	¥200
花子	¥100	¥200	¥300	¥400	¥200

太郎君がリッチな食事をとるとき、花子さんは貧乏な食事で我慢してるの？

	月	火	水	木	金	
太郎	¥400	¥300	¥100	¥200	¥200	¥240
偏差	160	60	-140	-40	-40	
花子	¥100	¥200	¥300	¥400	¥200	¥240
偏差	-140	-40	60	160	-40	
積	-22,400	-2,400	-8,400	-6,400	1,600	-7,600

平均

共分散

2変数の相関

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

□ 共分散 *covariance*

共分散って、一体何を測ってるの？

7

2変数の相関

$$\text{COV}_{xy} = \begin{cases} + \rightarrow \text{正の相関} \\ 0 \rightarrow \text{無相関} \\ - \rightarrow \text{負の相関} \end{cases}$$

□ 共分散 *covariance*

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

じゃあ、「相関の強さ」を「共分散の大きさ」で表せる？

2変数の相関

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

□ 共分散 *covariance*

□ 例: 文教太郎君と湘南花子さんの昼食費

	月	火	水	木	金
太郎	¥400	¥300	¥100	¥200	¥200
花子	¥100	¥200	¥300	¥400	¥200

太郎君がリッチな食事をとるとき、花子さんは貧乏な食事で我慢してるの？

□ 例: 文教次郎君と湘南花子さんの昼食費

	月	火	水	木	金
次郎	¥40万	¥30万	¥10万	¥20万	¥20万
花子	¥100	¥200	¥300	¥400	¥200

超リッチな食事をする次郎君と比べたら、花子さんの食事ってどうなの？

2変数の相関

□ 共分散 *covariance*

測定単位が変わると、相関の度合いが変わってしまう！

	月	火	水	木	金	
太郎	¥400	¥300	¥100	¥200	¥200	¥240
偏差	160	60	-140	-40	-40	
花子	¥100	¥200	¥300	¥400	¥200	¥240
偏差	-140	-40	60	160	-40	
積	-22,400	-2,400	-8,400	-6,400	1,600	-7,600

平均 →

共分散 →

	月	火	水	木	金	
次郎	¥40万	¥30万	¥10万	¥20万	¥20万	¥24万
偏差	16万	6万	-14万	-4万	-4万	
花子	¥100	¥200	¥300	¥400	¥200	¥240
偏差	-140	-40	60	160	-40	
積	-2,240万	-240万	-840万	-640万	160万	-760万

平均 →

共分散 →

2変数の相関

□ 相関係数 *correlation*

$$r_{xy} = \frac{COV_{xy}}{S_x \cdot S_y}$$
$$(-1 \leq r_{xy} \leq 1)$$

ピアソンの積率相関係数
Pearson's product-moment correlation coefficient

$$r_{xy} = \begin{cases} 1 & \text{正の相関} \\ 0 & \text{無相関} \\ -1 & \text{負の相関} \end{cases}$$

共分散をそれぞれのデータ x_i, y_i の標準偏差で割ることにより、測定単位を気にせずに、2種類のデータの関係の強さを表せる。

■ 注意

- 相関係数は、2つの変数の直線的関係を見るためのもの。曲線関係が認められる場合等には向かない
- 相関係数は、因果関係を保証するものではない。

2変数の相関

相関係数 correlation

測定単位が変わっても、相関の度合いは変わらない

	月	火	水	木	金	Ave.	St.Dev.
太郎	¥400	¥300	¥100	¥200	¥200	¥240	101.98
偏差	160	60	-140	-40	-40	Ave.	St.Dev.
花子	¥100	¥200	¥300	¥400	¥200	¥240	101.98
偏差	-140	-40	60	160	-40	Cov.	Corr.
積	-22,400	-2,400	-8,400	-6,400	1,600	-7,600	-0.731

	月	火	水	木	金	Ave.	St.Dev.
次郎	¥40万	¥30万	¥10万	¥20万	¥20万	¥24万	101,980
偏差	16万	6万	-14万	-4万	-4万	Ave.	St.Dev.
花子	¥100	¥200	¥300	¥400	¥200	¥240	101.98
偏差	-140	-40	60	160	-40	Cov.	Corr.
積	-2,240万	-240万	-840万	-640万	160万	-760万	-0.731

2変数の相関

順序尺度に対する相関係数

スピアマンの順位相関係数 Spearman rank correlation coefficient

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - Q_i)^2 \quad (-1 \leq r_s \leq 1)$$

順位が完全に一致しているとき $r_s = +1$
順位が完全に逆のとき $r_s = -1$

ケンドールの順位相関係数 Kendall tau rank correlation coefficient

$$r_K = \frac{G - H}{G + H} \quad (-1 \leq r_K \leq 1)$$

順位が完全に一致しているとき $r_K = +1$
順位が完全に逆のとき $r_K = -1$

10

相関関係

★順位相関係数を使うときは？
データが選好順位(順序尺度)で与えられている場合
 $A: R_1, R_2, \dots, R_n$

参考: その他の相関係数

□ 例題: 男女それぞれが好きな花の順番

	桜	菊	薔薇	梅	百合	鬱金香	カーネーション	椿
男	1	2	3	4	5	6	7	8
女	3	1	2	5	4	7	6	8

出展: (『統計学入門』p.55)

☆ (スピアマンの) 順位相関係数

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - Q_i)^2$$
$$= 1 - \frac{6}{8^3 - 8} \{ (1-3)^2 + (2-1)^2 + \dots + (8-8)^2 \}$$
$$= 1 - \frac{1}{84} \cdot 10 = \frac{37}{42} \approx 0.881$$

ピアソンの積率相関係数を順序尺度に素直にあてはめたもの

☆ (ケンドールの) 順位相関係数

	菊	薔	梅	百	鬱	カ	椿
桜	×	×	○	○	○	○	○
菊			○	○	○	○	○
薔				○	○	○	○
梅					×	○	○
百						○	○
鬱							×
カ							○

桜 v.s. 椿
★男: 1<8
★女: 3<8 } 正順

鬱 v.s. 力
★男: 6<7
★女: 7>6 } 逆順

G: 正順の数=24
H: 逆順の数=4

$$r_K = \frac{G - H}{n(n-1)/2} = \frac{24 - 4}{8(8-1)/2} = \frac{5}{7} \approx 0.714$$

全対 (n(n-1)/2個) について、正順と逆順の個数の差を比較したもの

演習5

相関係数を計算しよう

□ 右のデータ x, y について、

- それぞれの分散 S_x^2, S_y^2 を計算せよ。
- 共分散 cov_{xy} を計算せよ。
- (ピアソンの積率) 相関係数 r_{xy} を計算せよ。

x	1	3	5	7	9
y	4	6	2	0	3

□ 右のA君, Bさんの色の好みに関する選好順位データについて、

- (スピアマンの) 順位相関係数 r_s を計算せよ。
- (ケンドールの) 順位相関係数 r_K を計算せよ。

	赤	青	橙	緑	紫
A	1	2	3	4	5
B	4	5	2	1	3

最後に...

□ 統計解析・予測手法

記述統計学
descriptive statistics

度数分布, 代表値,
散らばり, 相関関係,
etc.

推測統計学
inferential statistics

確率分布,
母集団・標本,
推定, 検定, etc.

多変量解析
multivariate analysis

重回帰分析, 主成分分析,
判別分析, 数量化理論,
etc.

参考文献

- ✓ □ 東大教養統計教室編「統計学入門」東大出版会(1991)
- 東大教養統計教室編「自然科学の統計学」東大出版会(1992)
- ✓ □ 村上雅人「なるほど統計学」海鳴社(2002)
- ✓ □ 金子治平ほか「よくわかる統計学 I」ミネルヴァ書房(2007)
- ✓ □ 大村平「改訂版 統計解析のはなし」日科技連(2006,1980)
- 大村平「QC数学のはなし」日科技連(2003)
- ✓ □ 高橋信「マンガでわかる統計学」オーム社(2004)
- ✓ □ 田栗正章ほか「やさしい統計入門」講談社(2007)
- 桑田秀夫「経営・経済系のための統計学」日科技連(1992)
- J.アルバート&J.ベネット「メジャーリーグの数理科学」シュプリンガー(2004)
- 間瀬茂他「工学のためのデータサイエンス入門」数理工学社(2004)
- 荒木勉他「Excelで学ぶ統計解析」実教出版(2000)