2016年度秋学期　統計学　　第６回

データの関係を知る(1)—相関関係と因果関係

浅野　晃
関西大学総合情報学部

多変量データと多変量解析

変量とは

日本男性の身長は分布する

分布する量を［変量］という

統計学は，
分布している変量から情報を引き出す
手法

「多」変量とは

２つ以上の変量の組み合わせで
表現されるデータ

「入学試験の点数」←数学・英語・国語…

［多変量データ］
という

変量　変量　変量

多変量データを扱う統計学を
［多変量解析］という

多変量解析では

変量の間の関係が問題になる

たとえば

数学の点数の高い人は　英語の点数も高い

数学の点数の高い人は　国語の点数が低い

…という傾向にある

この傾向を見つけるのが，　[相関分析]

[回帰分析]

相関関係と散布図

A.Asano, Kansai Univ.

相関関係

２つの変量からなるデータを考える

さっきの

[正の相関関係]

数学の点数の高い人は　英語の点数も高い

数学の点数の高い人は　国語の点数が低い

[負の相関関係]

という傾向にある

変量どうしの互いの増減の傾向

[相関関係]

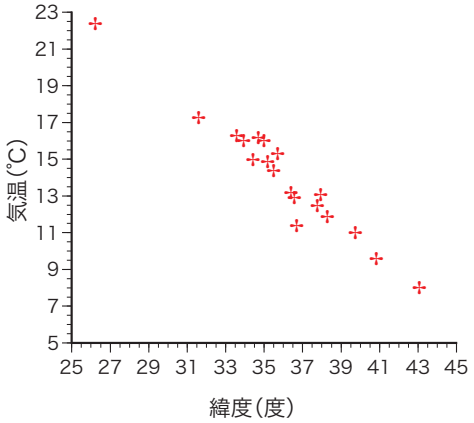
A.Asano, Kansai Univ.
2016年度秋学期　統計学

散布図

多変量データを目に見えるように描く

地名	緯度（度）	気温（℃）
札幌	43.05	8.0
青森	40.82	9.6
秋田	39.72	11.0
仙台	38.27	11.9
福島	37.75	12.5
宇都宮	36.55	12.9
水戸	36.38	13.2
東京	35.68	15.3
新潟	37.92	13.1
長野	36.67	11.4
静岡	34.97	16.0
名古屋	35.17	14.9
大阪	34.68	16.2
鳥取	35.48	14.4
広島	34.40	15.0
高知	33.55	16.3
福岡	33.92	16.0
鹿児島	31.57	17.3
那覇	26.20	22.0

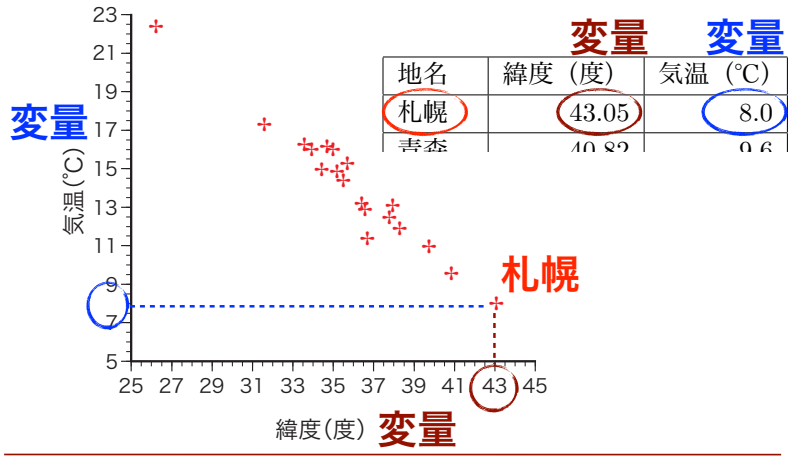
表 1: 日本の都市の緯度と気温



A.Asano, Kansai Univ.
2016年度秋学期　統計学

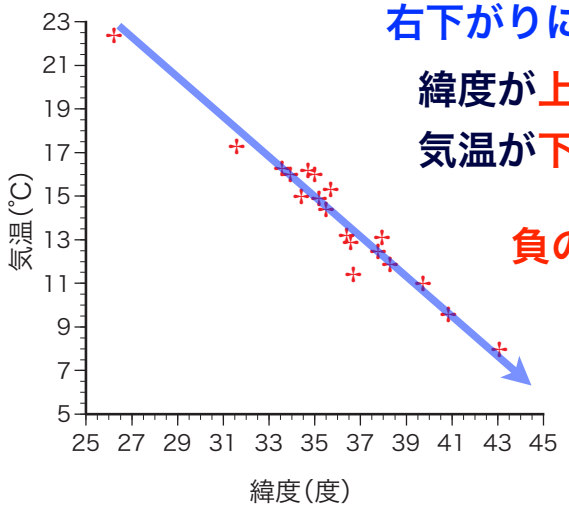
散布図

多変量データを目に見えるように描く



地名	緯度（度）	気温（℃）
札幌	43.05	8.0
青森	40.82	9.6

散布図と相関関係



右下がりに並ぶ

緯度が上がると

気温が下がる傾向

負の相関関係

相関の強弱

参考資料の散布図（47都道府県について）

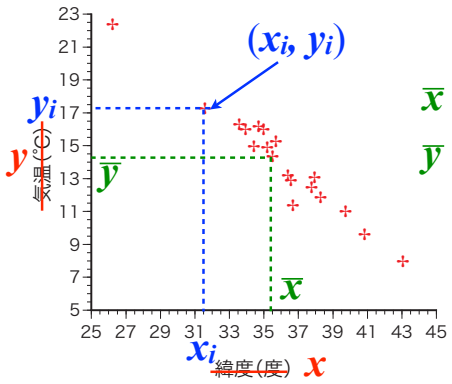
「統計学入門」（東京大学出版会）
４４ページの図（さまざまな散布図の例）を示して，
相関の強弱や無相関について，
スライド2枚にわたって説明しました。

共分散と相関係数

相関係数

相関の正負・強弱を数字で表す

ここからは，緯度・気温ではなく一般的に



$\bar{x}$　x だけの平均

$\bar{y}$　y だけの平均

相関係数

x，y の［共分散］

［相関係数］

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}}$$

x の偏差　y の偏差

x の平均　y の標準偏差

x の偏差

x の分散

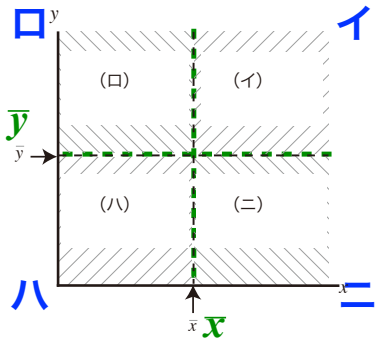
x の標準偏差

（nはデータサイズ）

共分散の意味

x，y の共分散　$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n$

x の偏差　y の偏差



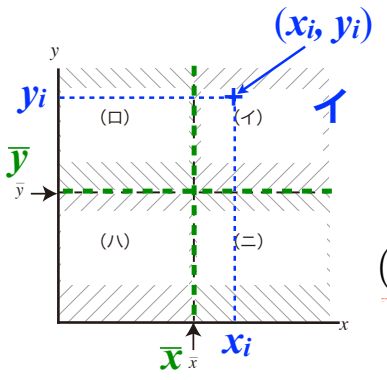
イ・ロ・ハ・ニで

$$(x_i - \bar{x})(y_i - \bar{y})$$

の値はどうなる？

共分散の意味

(x_i, y_i) が「イ」の領域にあるとすると



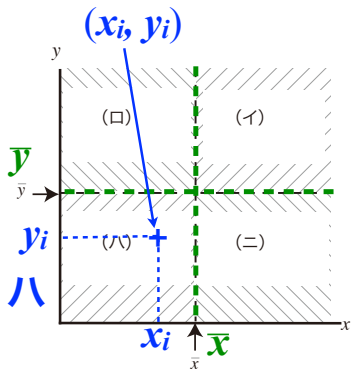
$$x_i - \bar{x} > 0$$

$$y_i - \bar{y} > 0$$

$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$

共分散の意味

(x_i, y_i) が「ハ」の領域にあるとすると



$$x_i - \bar{x} < 0$$

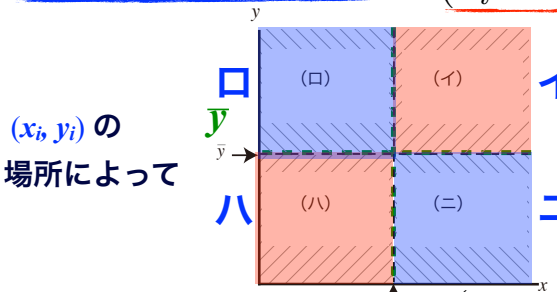
$$y_i - \bar{y} < 0$$

$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$

共分散の意味

$(x_i - \bar{x})(y_i - \bar{y}) < 0$

$(x_i - \bar{x})(y_i - \bar{y}) > 0$



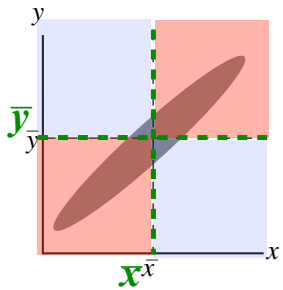
$(x_i - \bar{x})(y_i - \bar{y}) > 0$

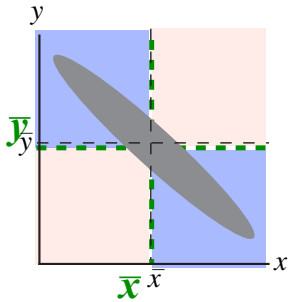
$(x_i - \bar{x})(y_i - \bar{y}) < 0$

(x_i, y_i) が $(\bar{x}, \bar{y})$から離れているほど，
絶対値が大きくなる

共分散の意味

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n$ は



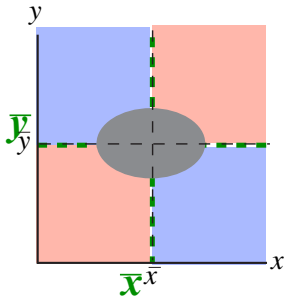


正で大きな値
→強い正の相関

負で絶対値が大きい
→強い負の相関

共分散の意味

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n$ は

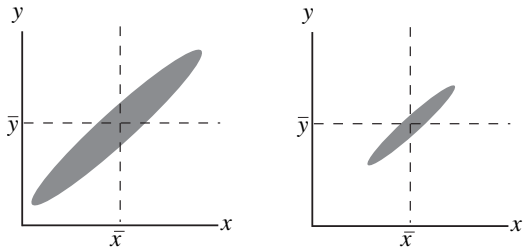


差し引きゼロ
→無相関

共分散と相関係数

相関係数＝共分散

÷ (xの標準偏差 × yの標準偏差)



これらの相関の強さは同じ
→標準偏差で割って調整する

相関係数は
-1〜0〜1

ちょっと問題

問題１

国民所得と酒の消費量の間には正の相関がある。だから，国民が酒をたくさん飲めば所得が増える。

相関関係と因果関係は異なる。

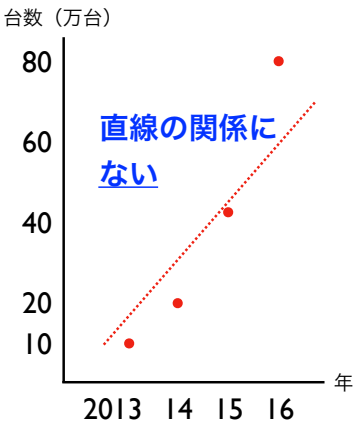
問題２

ある電気製品の普及台数は，発売以来毎年倍に増えている。発売後の年数と普及台数の相関係数は，非常に強い相関であるから，ほぼ１である。

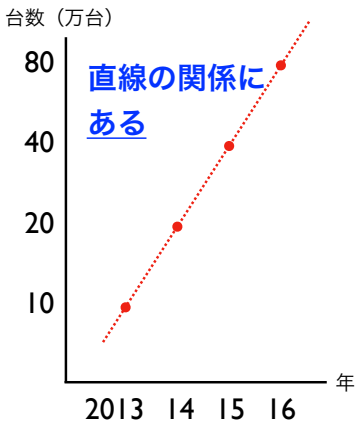
直線状の関係ではないから，
相関係数が１にはならない

問題２

「毎年倍になっている」



対数目盛りに変える
（1目盛＝「2倍」）



みかけ上の相関

A.Asano, Kansai Univ.

みかけ上の相関

小学生については，身体が大きいと
試験の成績が良い

？？？

全学年の児童に同じ問題で試験をすれば。

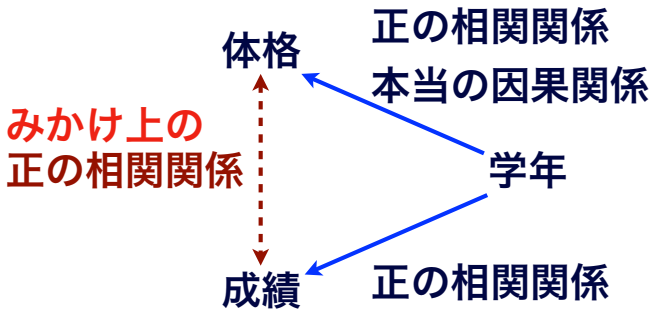
「体格」と「成績」には正の相関関係

なぜ？

A.Asano, Kansai Univ.
2016年度秋学期　統計学

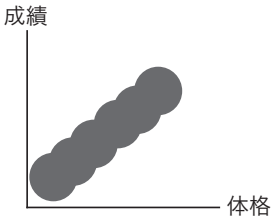
みかけ上の相関

なぜ？



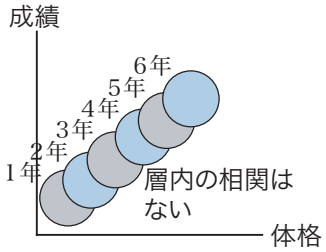
A.Asano, Kansai Univ.
2016年度秋学期　統計学

層別



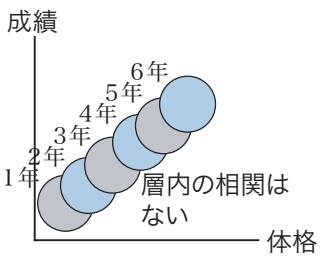
正の相関関係

実は

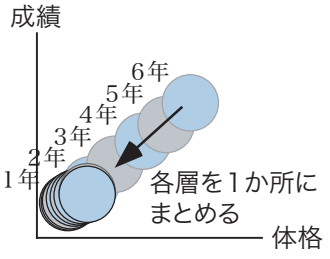


内部に「学年」の
層がある

層別



内部に「学年」の
層がある

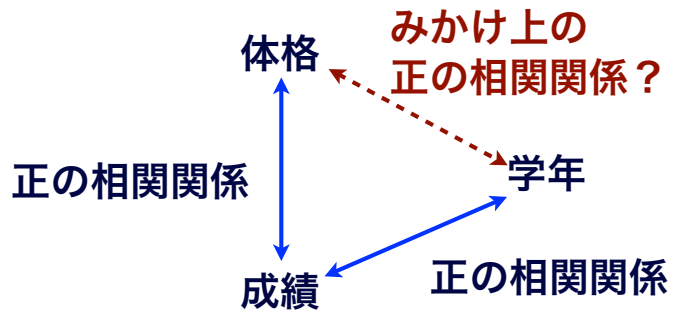


層に分けて，
ひとつにまとめる

学年の影響を除いた［偏相関係数］

ところで

こうはならないの？



統計学の上では，こう考えても同じ
ならないのは，統計学以外の知識による