2022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

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(この例なら身長)を[変量]という

統計学とは,
「分布している変量から情報を引き出す手法」ということができる

「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を,
[多変量データ]という

例えば「入学試験の点数」← 数学・英語・国語…

こういうデータが 多変量データ　　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

…という傾向にあるか？

※数学の点数の高い人は「必ず」英語の点数も高いのか？　のではなくて，
数学の点数の高い人は「多くの場合」英語の点数も高いのか？　ということ

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🤔

相関関係

2つの変量からなる多変量データを考える

「数学の点数の高い人は　英語の点数も高い」傾向にある

この場合，[正の相関関係]があるという

「数学の点数の高い人は　国語の点数が低い」傾向にある

この場合，[負の相関関係]があるという

変量どうしの互いの増減の傾向を[相関関係]という

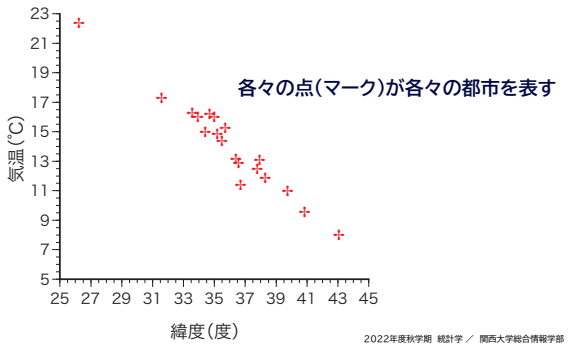
散布図(scattergram)

多変量データを目に見えるように描く方法

「各都市の緯度と気温」という多変量データを，散布図に描いてみる。

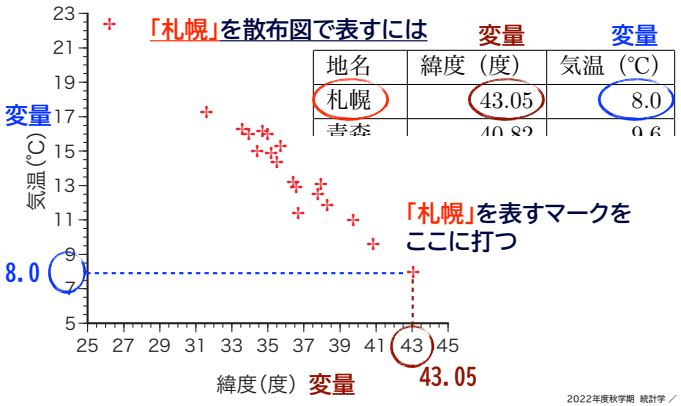
地名	緯度（度）	気温（℃）
札幌	43.05	8.0
青森	40.82	9.6
秋田	39.72	11.0
仙台	38.27	11.9
福島	37.75	12.5
宇都宮	36.55	12.9
水戸	36.38	13.2
東京	35.68	15.3
新潟	37.92	13.1
長野	36.67	11.4
静岡	34.97	16.0
名古屋	35.17	14.9
大阪	34.68	16.2
鳥取	35.48	14.4
広島	34.40	15.0
高知	33.55	16.3
福岡	33.92	16.0
鹿児島	31.57	17.3
那覇	26.20	22.0

表 1: 日本の都市の緯度と気温

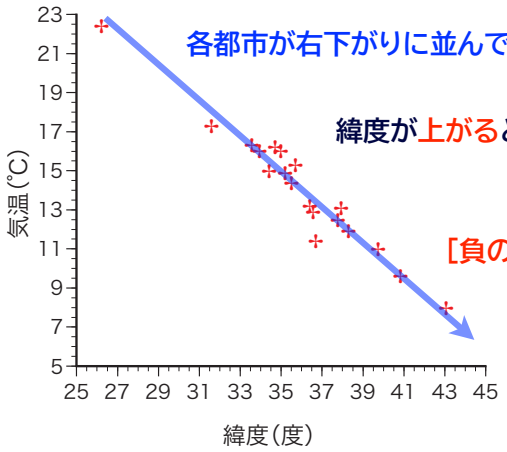


散布図の描き方

多変量データを目に見えるように描く

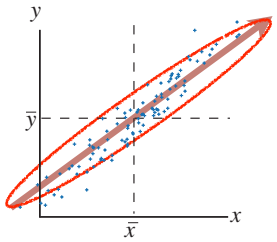


散布図と相関関係



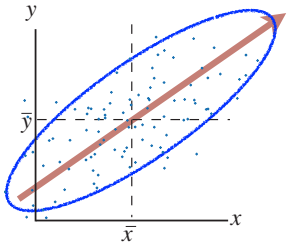
相関の強弱

どちらも正の相関関係だが，強い相関と弱い相関がある



ほぼ完全に，一直線上に並んでいる

［強い相関］

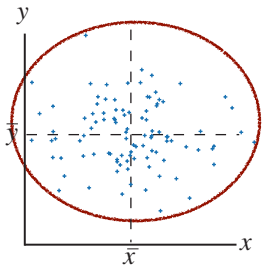


直線に沿って並んではいるが，
まわりに広がっている

［弱い相関］

相関の強弱

右上がりでも右下がりでもない場合は



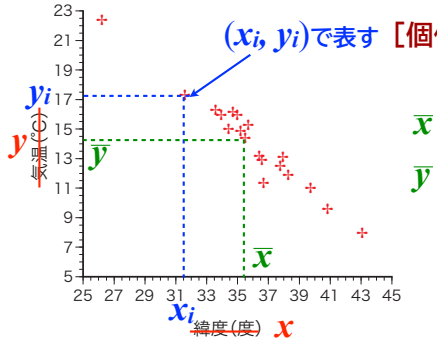
［相関関係なし］

［無相関］

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



相関係数

相関の正負・強弱を数字で表す

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

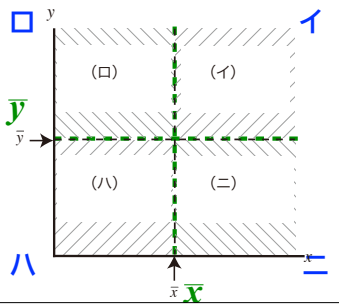
（nはデータサイズ）

共分散の意味

x, y の共分散

x の偏差　y の偏差

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



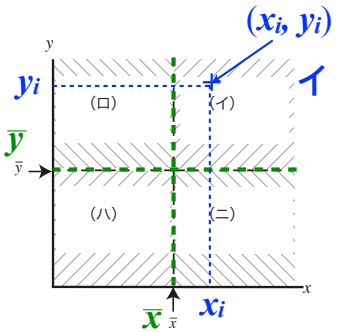
イ・ロ・ハ・ニで

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

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



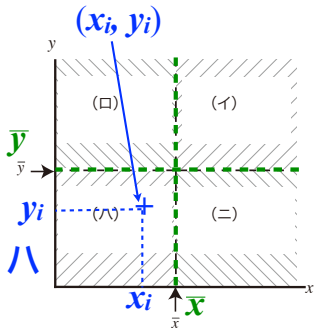
$$x_i - \bar{x} > 0$$

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

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

共分散の意味

(x_i, y_i) が「ハ」の領域にあるとすると



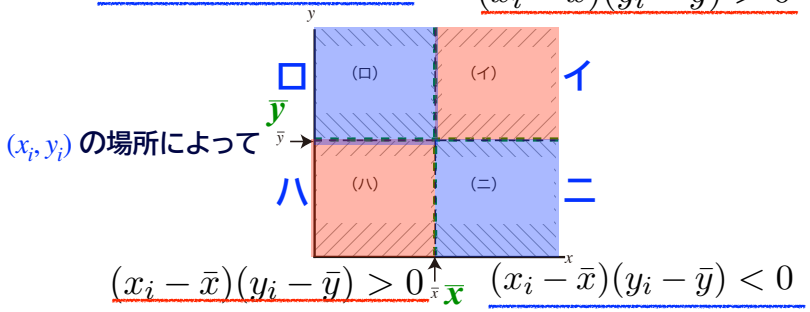
$$x_i - \bar{x} < 0$$

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

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

共分散の意味

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

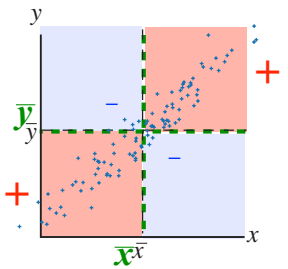


$(x_i - \bar{x})(y_i - \bar{y}) > 0$ $(x_i - \bar{x})(y_i - \bar{y}) < 0$

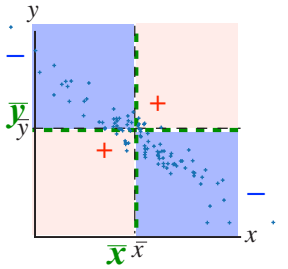
(x_i, y_i) が $(\bar{x}, \bar{y})$から離れているほど，絶対値が大きくなる

共分散の意味

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



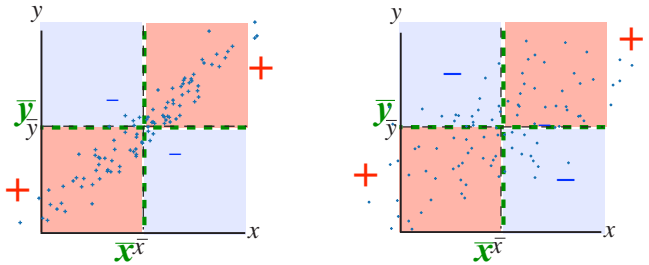
正で大きな値
→強い正の相関？



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

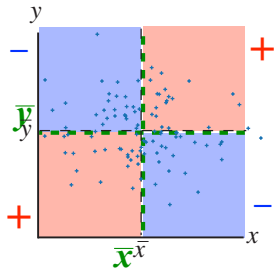


正で大きな値
→強い正の相関？

正だが大きくない
→弱い正の相関？

共分散の意味

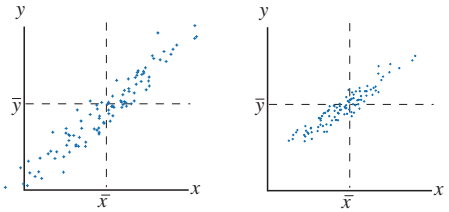
$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})/n$ は



差し引きゼロ
→無相関

さっき？がついていたのは一共分散と相関係数

相関係数＝共分散 ÷（x の標準偏差 × y の標準偏差）



相関係数は
-1～0～1の
範囲にある

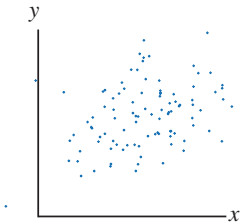
左のほうが，x,y ともよりばらついているので，共分散は左のほうが大きい

しかし，これらの相関の強さは同じ
→ そうなるように，標準偏差で割って調整したのが，相関係数

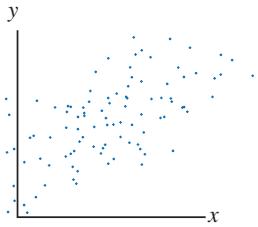
中くらいの相関

相関係数が-1～0～1の範囲なら，相関係数0.5は中くらいの相関？

相関係数 0.5



相関係数 0.7



こうである理由，
また「中くらい」の意味は，
次回説明します

こちらが「中くらいの相関」

テキストでは

こんな表で，長野～鹿児島の緯度と気温の相関係数を求めています

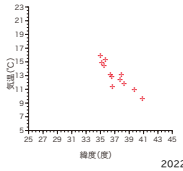
表 2: 相関係数を求める

地名	緯度（度）	気温（℃）	緯度の偏差	左の２乗	気温の偏差	左の２乗	両偏差の積
長野	36.67	11.4	2.18	4.752	−3.878	15.037	−8.454
静岡	34.97	16.0	0.48	0.230	0.722	0.522	0.347
名古屋	35.17	14.9	0.68	0.462	−0.378	0.143	−0.257
大阪	34.68	16.2	0.19	0.036	0.922	0.850	0.175
鳥取	35.48	14.4	0.99	0.980	−0.878	0.770	−0.869
広島	34.40	15.0	−0.09	0.008	−0.278	0.077	0.025
高知	33.55	16.3	−0.94	0.884	1.022	1.045	−0.961
福岡	33.92	16.0	−0.57	0.325	0.722	0.522	−0.412
鹿児島	31.57	17.3	−2.92	8.526	2.022	4.089	−5.905
	緯度の平均 = 34.49	気温の平均 = 15.278		緯度の分散 = 1.800		気温の分散 = 2.562	共分散 = −1.813
							相関係数 = −0.844

相関係数 -0.844

現在では，こういう手計算を行うことは
まずありません。ただ，

数字を入れるだけで計算できてしまうと，
大失敗をする可能性がある



散布図を見て，妥当な数値か
どうか確かめましょう。

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，別の概念である。

2つの変量の間に相関関係がある，といっても，
それはどちらがどちらの原因であるともいっていないし，
両者の間に因果関係があるかないかもいっていない。

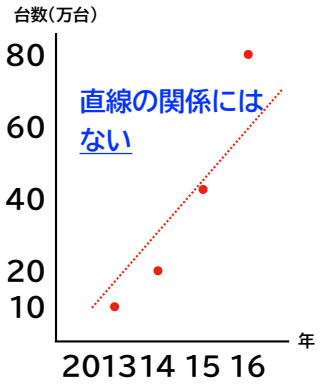
問題2

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。

対数目盛り

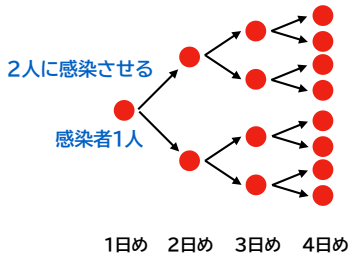
「毎年倍になっている」



縦軸を1目盛で「2倍」を表す
対数目盛りに変える

感染症の感染者数のグラフ

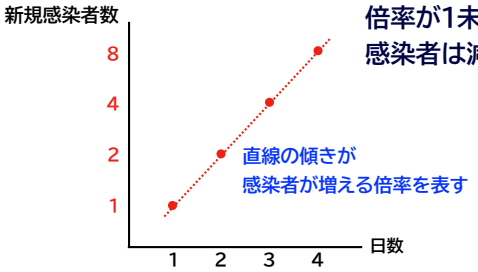
対数目盛りは，感染者数のグラフに用いられている
感染では，「倍率」が問題になるため。



たとえば，
それぞれの感染者が1日に2人に感染させると，
新たに感染した人（新規感染者）の数は
毎日2倍になる

感染者数のグラフ

対数目盛りは，感染者数のグラフに用いられている
たとえば，それぞれの感染者が1日に2人に感染させると，
新たに感染した人の数は毎日2倍になる
対数目盛で直線の関係になる



倍率が1未満になると，直線は右下がりになり，
感染者は減っていく

※ウェブサイトの【参考リンク】にある
“Coronavirus tracked”では，
縦軸を通常の線形目盛（linear）と
対数目盛（logarithmic）に切り替
えることができます。

みかけ上の相関🤔

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

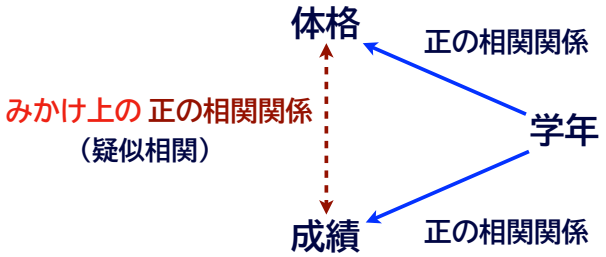
？？？

全学年の児童に同じ問題で試験をすれば。

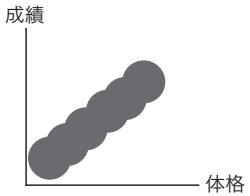
しかし，たしかに
「体格」と「成績」には正の相関関係　どうなってるの？

みかけ上の相関

どうなってるの？

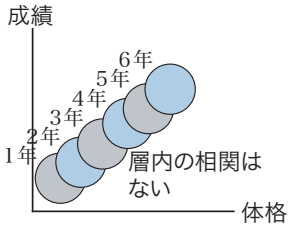


層別



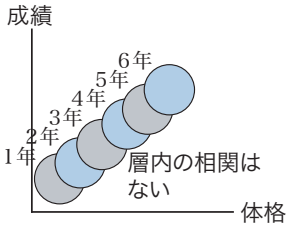
正の相関関係

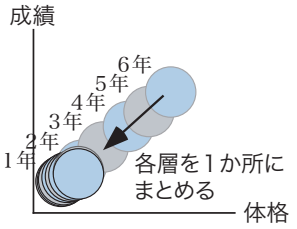
実は



内部に「学年」の層がある

層別



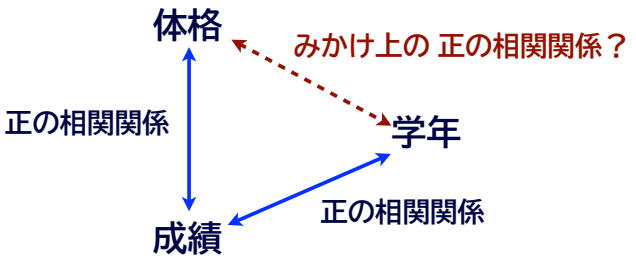


内部に「学年」の層がある　層に分けて，ひとつにまとめる

このようにして学年の影響を除いたのが[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，別の知識による

「みかけ上の」とは？

「嘘の相関」ではない。 相関関係は本当にある

「この現象が起きているしくみ」を
考えないなら，
(小学生の)体格から成績を
推測できる