データ間の関係を分析

広島大学 AI・データイノベーション教育研究センター
稲垣知宏

目標

データとデータの間にある関係を読み取り、定量的に扱えるようになる

この授業で紹介すること

- データ間の関係と共分散
- 相関係数の意味

キーワード

共分散、相関係数

こんなことはありませんか？

Aさんは、プログラミングが得意な人は他の科目も得意なのかどうか、あるいは全く関係ないのか、テスト結果のデータに基づいて分析したいと思いました。

テスト結果のデータから、どのような情報が得られるのでしょうか。

データ間の関係

まず、異なるデータの比較について考えます。

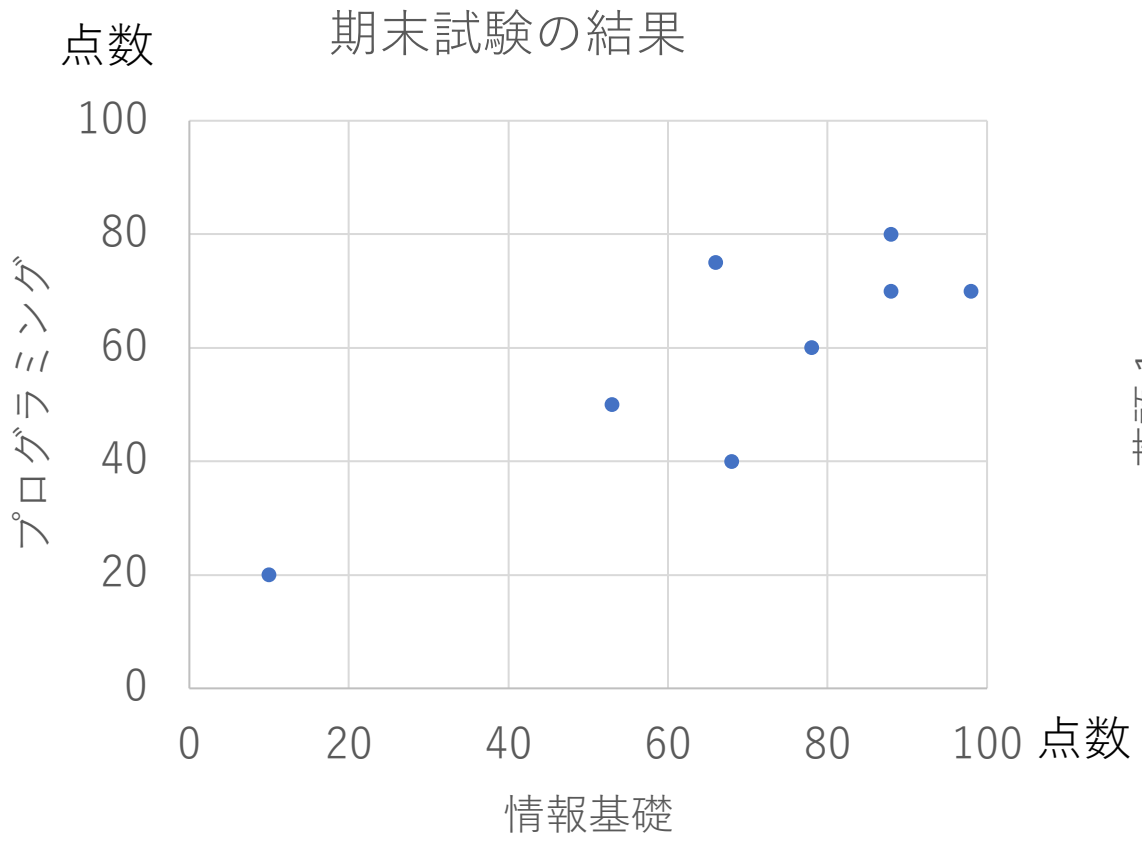
右表にある３科目の期末テスト結果を比較してみましょう。

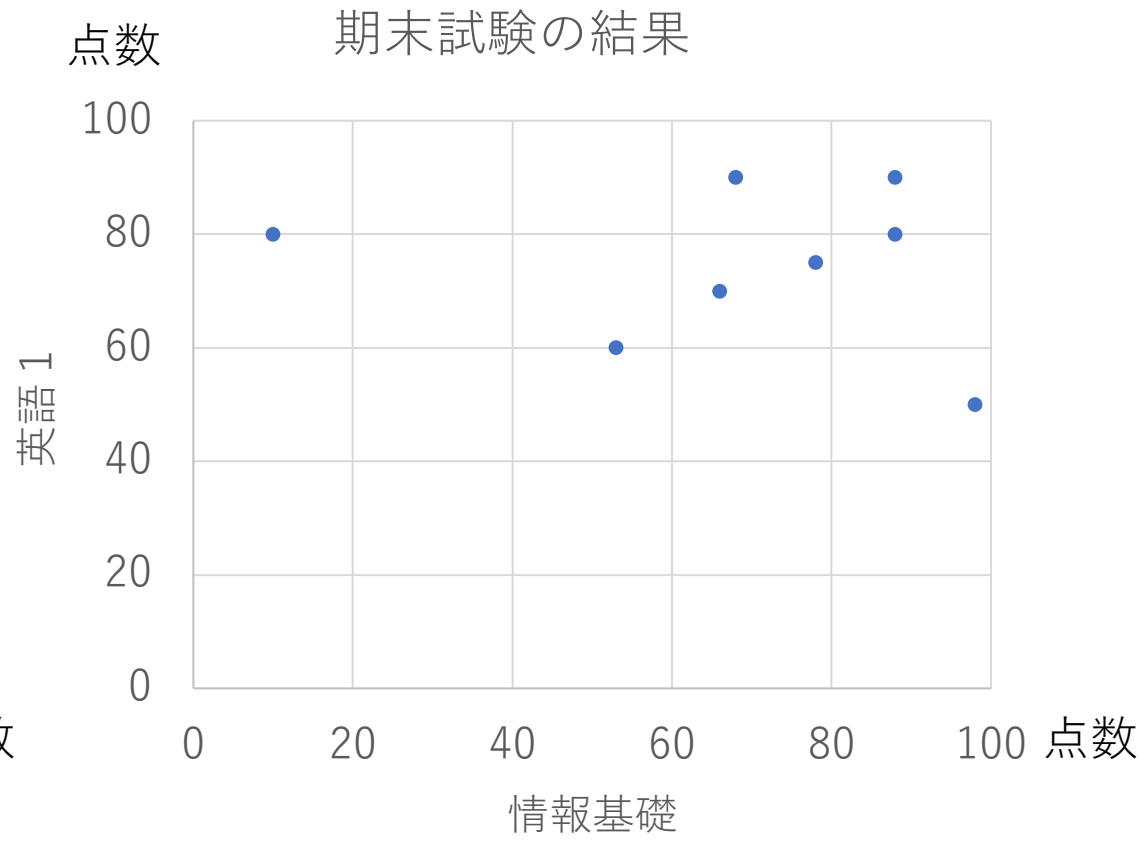
ばらつきの平均（平均値からのずれの２乗の平均）を分散と呼びます。

期末テストの結果

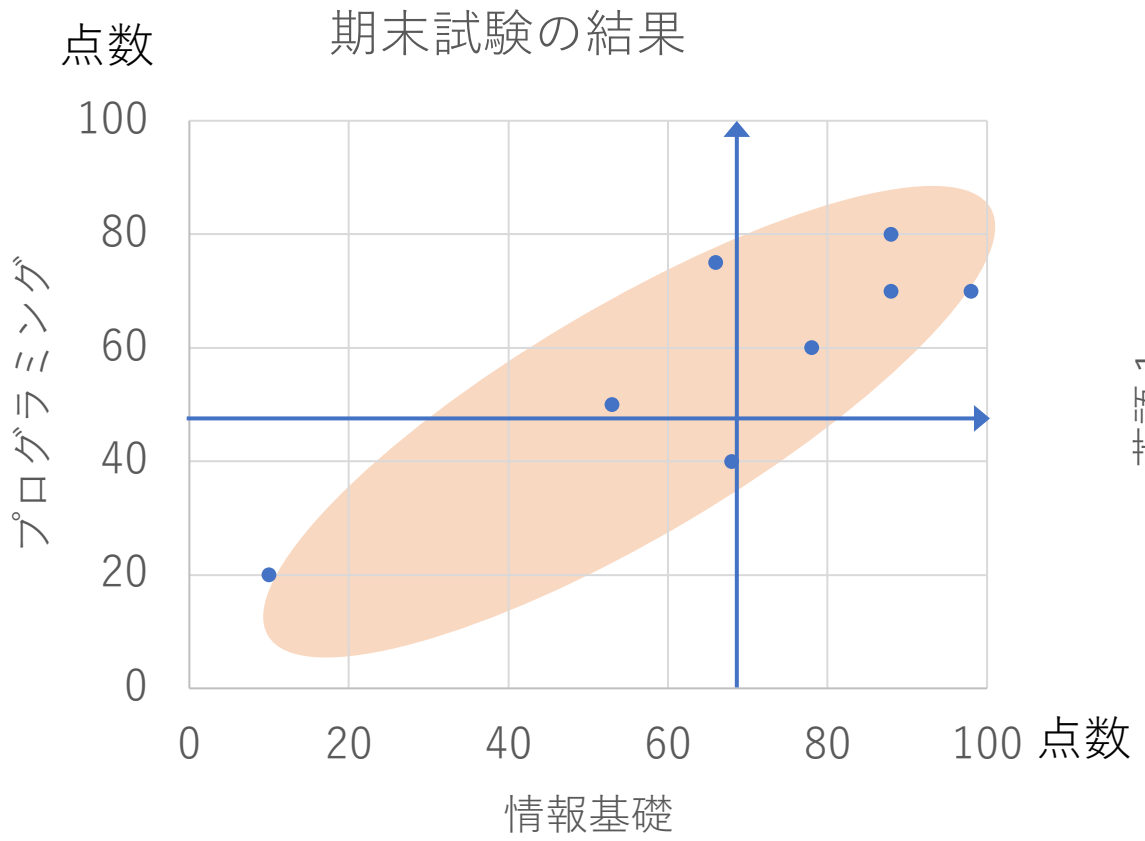
	情報基礎	プログラミング	英語１
A	98	70	50
B	88	80	80
C	68	40	90
D	53	50	60
E	10	20	80
F	66	75	70
G	78	60	75
H	88	70	90
平均	68.625	58.125	74.375
分散	673.734375	362.109375	171.484375

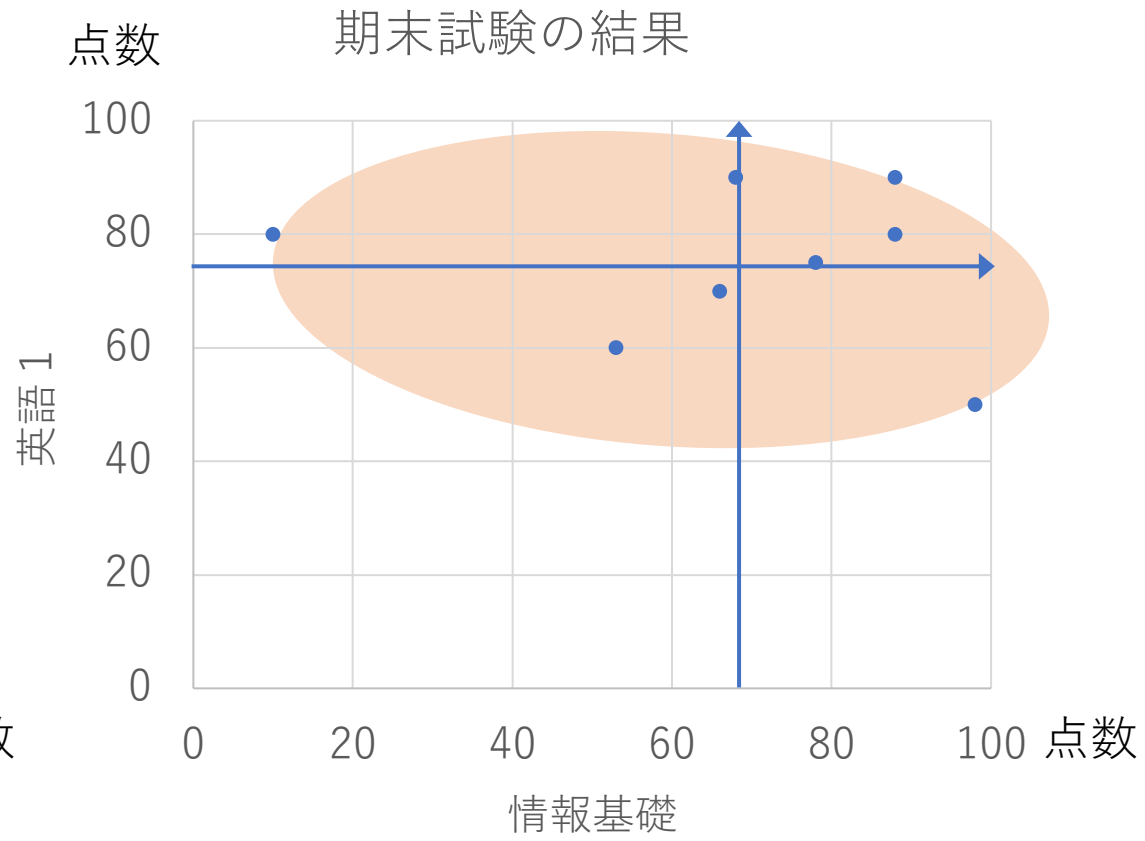
グラフ（散布図）で確認


期末試験の結果
点数
プログラミング
100
80
60
40
20
0
0
20
40
60
80
100
点数
情報基礎


期末試験の結果
点数
英語１
100
80
60
40
20
0
0
20
40
60
80
100
点数
情報基礎

グラフ（散布図）で確認


期末試験の結果
点数
プログラミング
100
80
60
40
20
0
0
20
40
60
80
100
点数
情報基礎


期末試験の結果
点数
英語１
100
80
60
40
20
0
0
20
40
60
80
100
点数
情報基礎

共分散

散布図の各点について縦軸方向、横軸方向の平均値からのズレの積を求めます。Aさんの場合、情報基礎とプログラミングのテスト結果については、

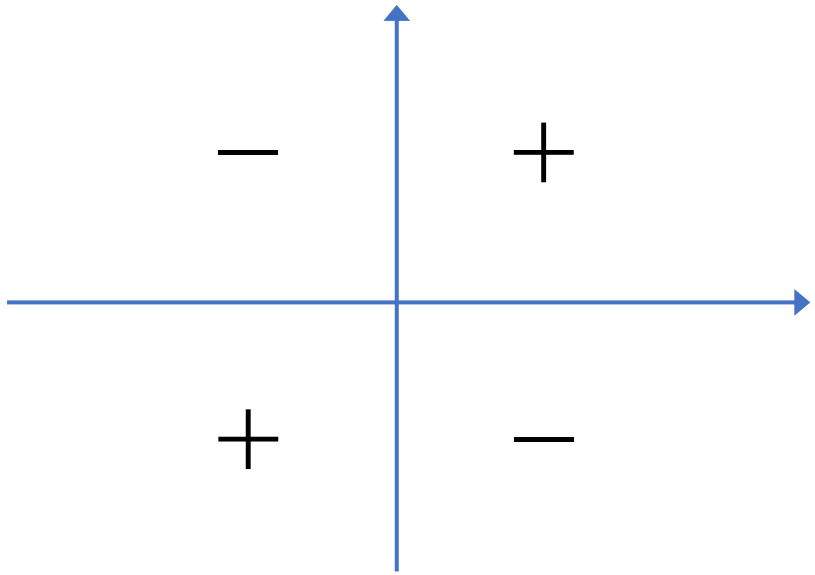
$$(98 - 68.625)(70 - 58.125) = 348.8...$$

となります。

散布図の右上と左下の点ではプラス、

左上と右下ではマイナスの値です。

この積の平均値を共分散と呼びます。



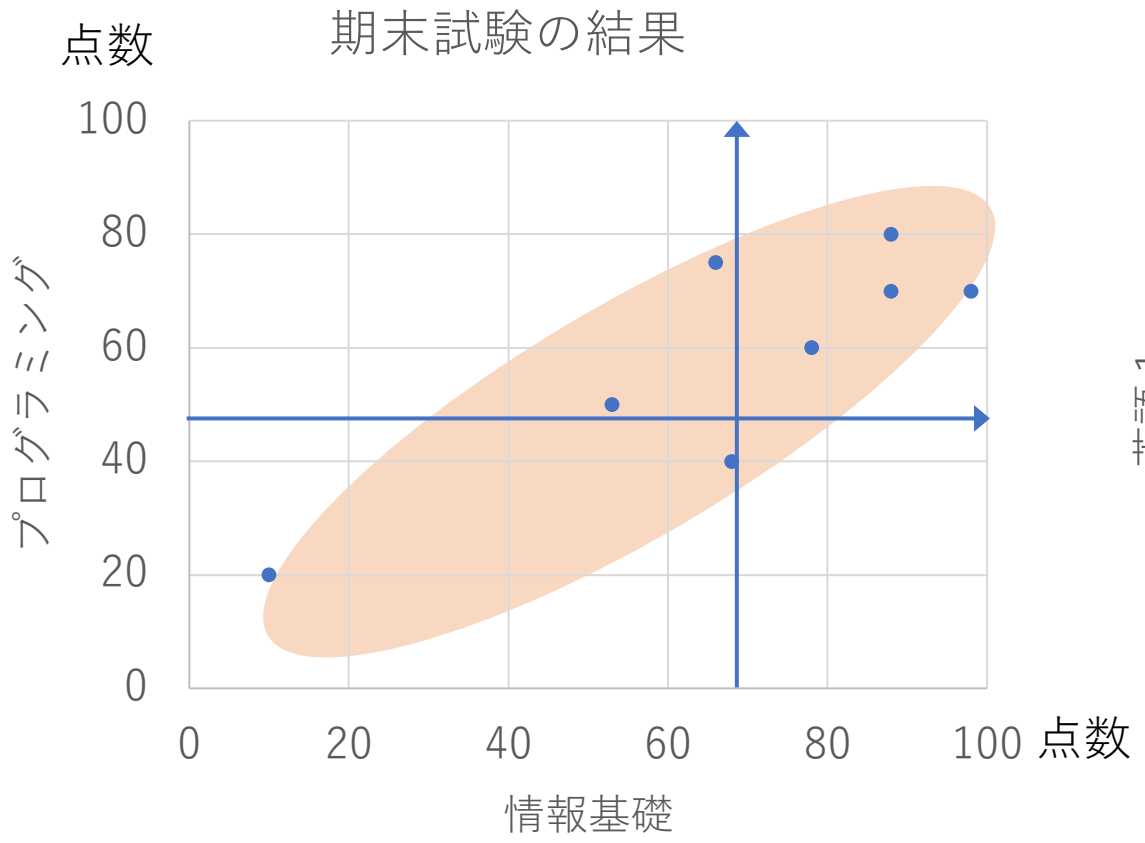
共分散

テスト結果の各組について、共分散を求めると

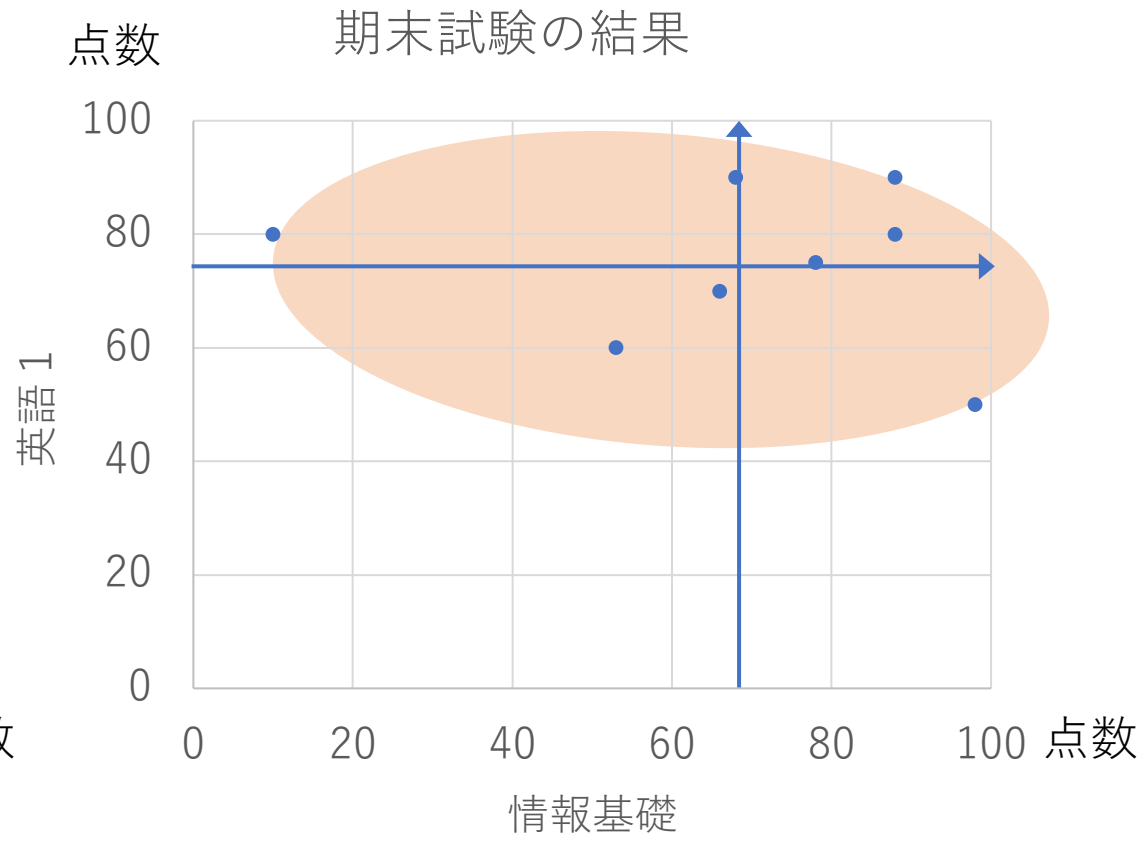
	情報基礎	プログラミング	英語１
情報基礎		418.671875	-50.234375
プログラミング	418.671875		-54.296875
英語１	-50.234375	-54.296875	

となり、情報基礎とプログラミングでは共分散がプラスになり、他はマイナスになっていることが分かります。

グラフ（散布図）で確認



共分散がプラス



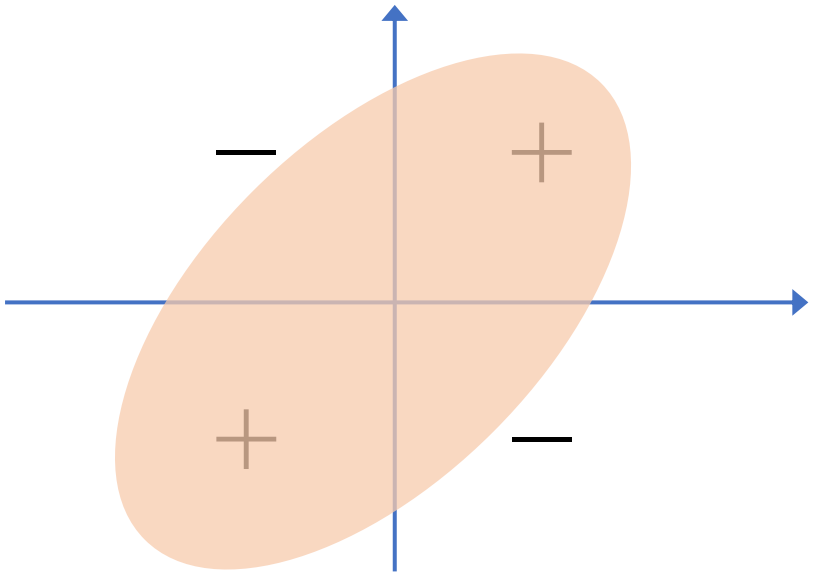
共分散がマイナス

例題

1. 共分散がプラスになった情報基礎とプログララミングの期末テストの結果の間には、どのような関係があると考えられますか？
2. 共分散がマイナスになった情報基礎と英語1の期末テストの結果の間の関係についてはどうですか？

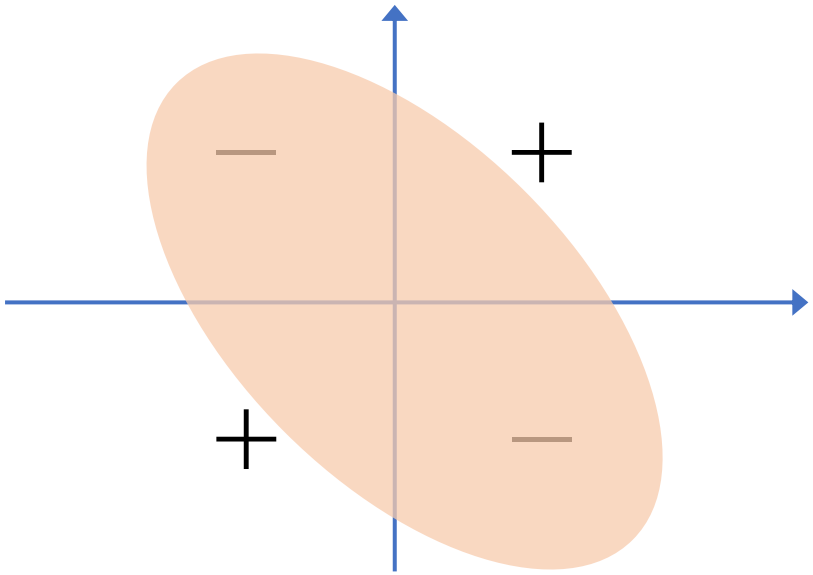
解説１：共分散がプラスの場合

- 共分散がプラスになっているデータ間では、一方が増えるともう一方も増えるという傾向があると考えられます。
- 例題の場合、情報基礎のテスト結果が高い人の方がプログラミングのテストの結果も高かい傾向があると考えられます。



解説：共分散がマイナスの場合

- 共分散がマイナスのデータ間では、一方が増えるともう一方は減るという傾向があると考えられます。
- 例題の場合、情報基礎のテスト結果が高い人の方は英語１のテストの結果が低い傾向があると考えられそうですが、この傾向は明らかではありません。



共分散と相関係数

相関があるかないかを確認するには、共分散を各データの分散の平方根（標準偏差）で割って、相関係数を求めて判断します。

	情報基礎	プログラミング	英語１
情報基礎		0.847636305	-0.14778977
プログラミング	0.847636305		-0.217892696
英語１	-0.14778977	-0.217892696	

相関係数は-1と1の間の値を取り、この値が1に近ければ、強い正の相関、-1に近ければ強い負の相関があります。

相関係数の意味

相関係数がどのくらいであれば相関があると言えるのかは考えている問題にも寄りますが、概ね、下表のように判断します。

相関係数	
0.7～1.0	強い正の相関
0.5～0.7	正の相関
-0.5～0.5	相関なし
-0.7～-0.7	負の相関
-1.0～-0.7	強い負の相関

なお、相関は因果関係があることと意味するわけではありません。

問題

右表の４科目の期末試験の結果について、英語１と英語２の間の相関係数を求め、情報基礎とプログラミングの間の相関とどちらが強いか議論しなさい。

	情報基礎	プログラミング	英語１	英語２
A	98	70	50	40
B	88	80	80	60
C	68	40	90	80
D	53	50	60	50
E	10	20	80	85
F	66	75	70	90
G	78	60	75	60
H	88	70	90	80
平均	68.625	58.125	74.375	68.125
分散	673.734375	362.109375	171.484375	287.109375