

富士未来学Ⅱ

ーデータ分析講座Ⅱ

データ分析講座Ⅱでできるようになること

散布図から相関関係の有無を読み取ることで、2変数の間の関係を記述することができる。相関係数を計算したり、回帰直線を求めたりすることで、2変量データの特徴を記述することができる。

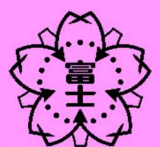
データ分析講座Ⅱで学ぶこと

散布図の作成の仕方、共分散と相関係数の計算方法とその意味を学ぶ。回帰直線の求め方など、2変量データの特徴を記述する方法を学ぶ。

月 日 ()

東京都立富士高等学校

東京都立富士高等学校附属中学校



中学2年 組 番 氏名



データ分析講座Ⅱ①

ループブックを見返しながら、自己の到達度を確認し、改善を図りましょう。

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
挑戦力	試行錯誤	主体的に学習に取り組む態度	1と2の記述	1と2において、1で散布図を作成して、2で相関関係について記述している。	1と2において、1で散布図を作成していない、または、2で相関関係について記述していない。	1と2において、1で散布図を作成していない、かつ、2で相関関係について記述していない。		
理数的解決力	分析解析	思考・判断・表現	4の記述	4において、正の相関関係と共分散の符号の関係を、①から③を記述し、④は理由とともに記述している。	4において、①から点の位置と偏差の符号の関係をみだして②を記述しているが、③や④を記述していない。	4において、①で偏差の計算や符号の判断ができず、表を埋めていない。		

1 散布図について

(1) 散布図

散布図は、2変量データ（二つの要素からなる1組のデータ）において、二つの要素の間の関係を視覚化し、分布の傾向をつかむための図です。

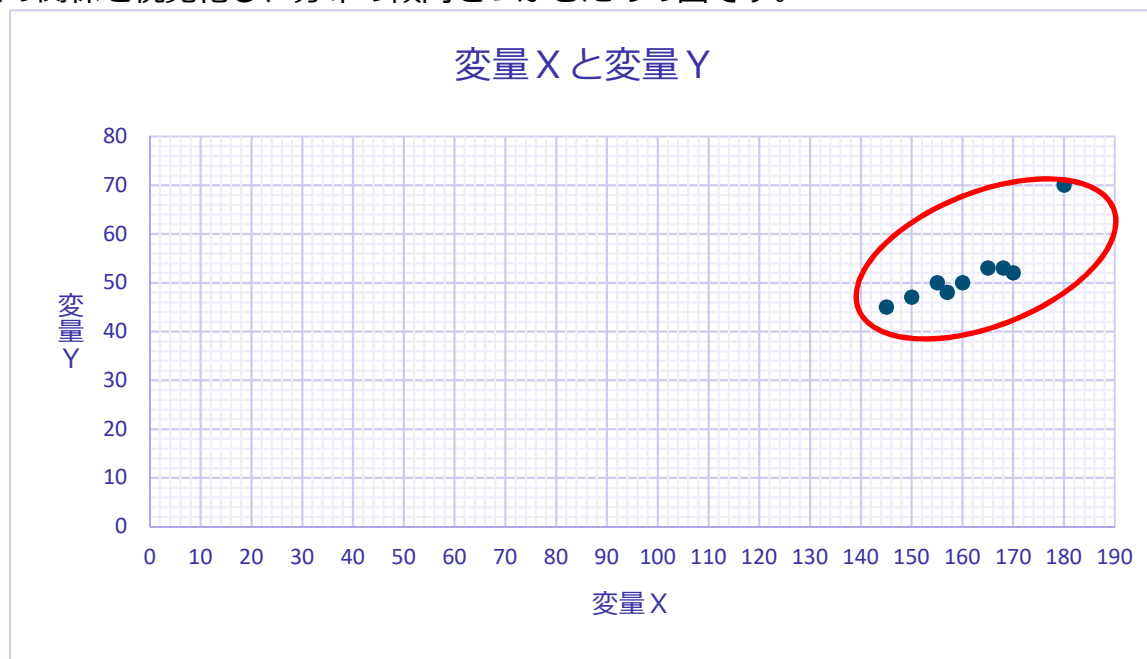


図1 変量Xと変量Yの散布図

図1の散布図からわかることは何でしょうか。



(2) 散布図の注意点

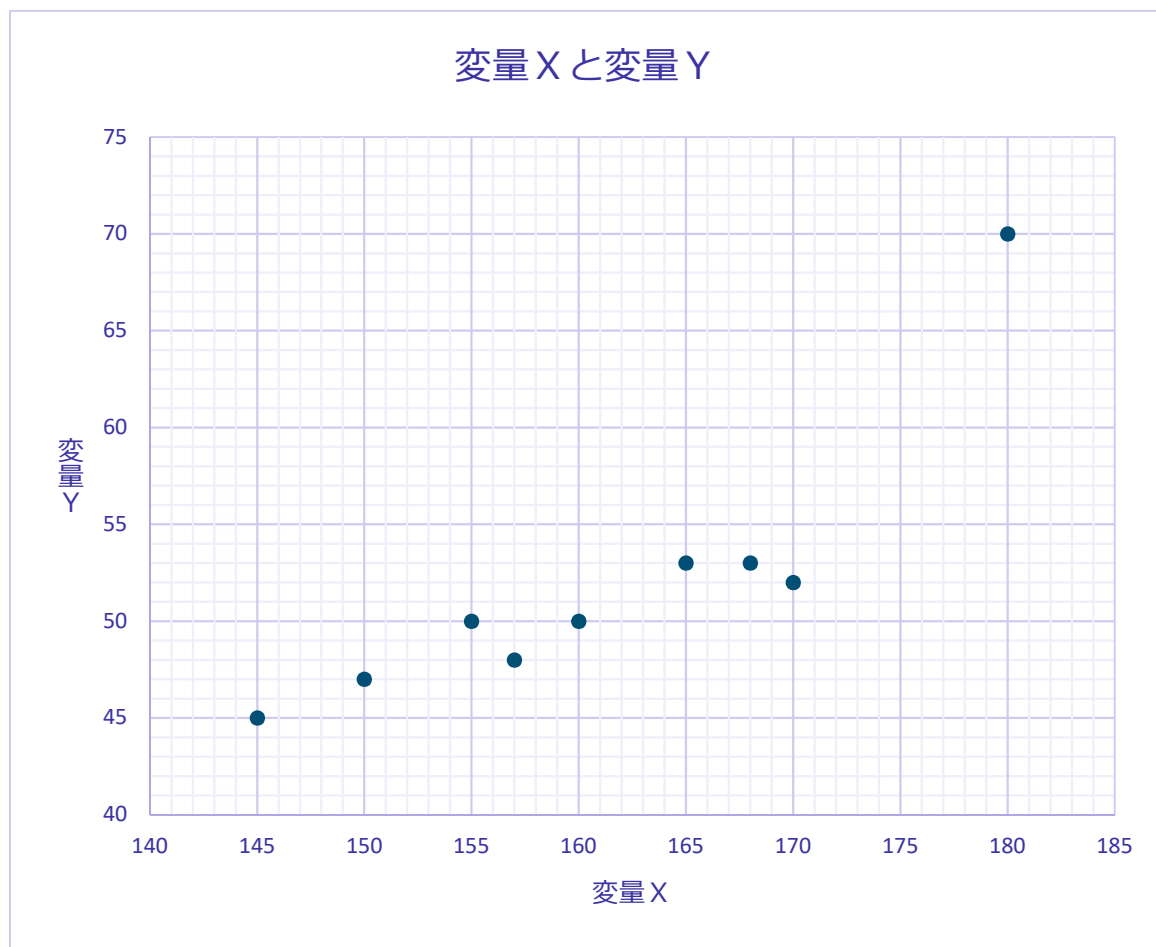


図2 変量 X と変量 Y の散布図

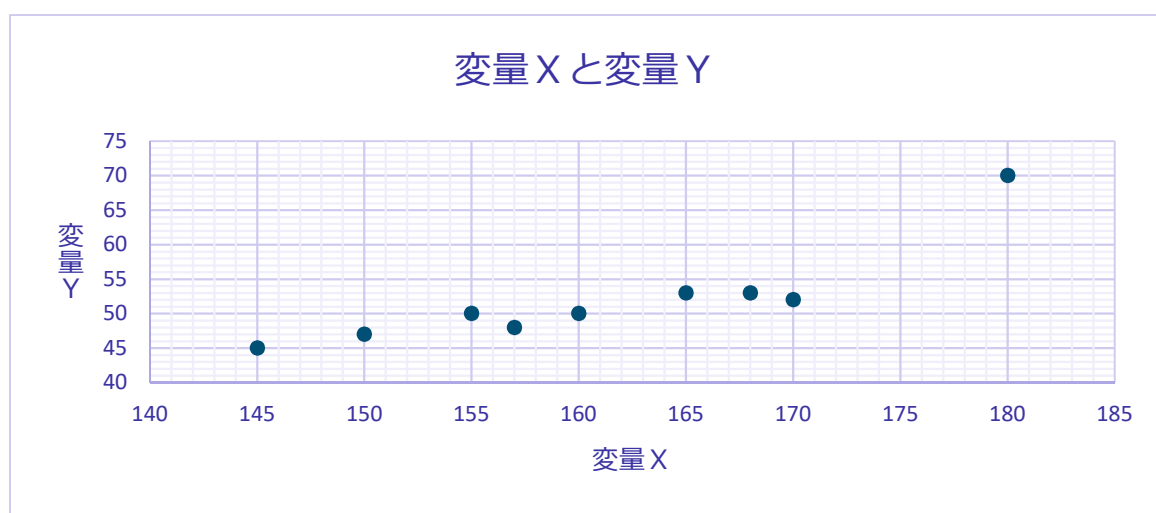


図3 変量 X と変量 Y の散布図

図2 と図3 において、散布図の傾向を読み取る際に、注意する点は何ですか。



図2と図3のように、メモリの幅や軸の最大値と最小値などをそれぞれ変えてみると、見え方が変わります。

(3) 散布図の作成

表1 数学と物理の得点

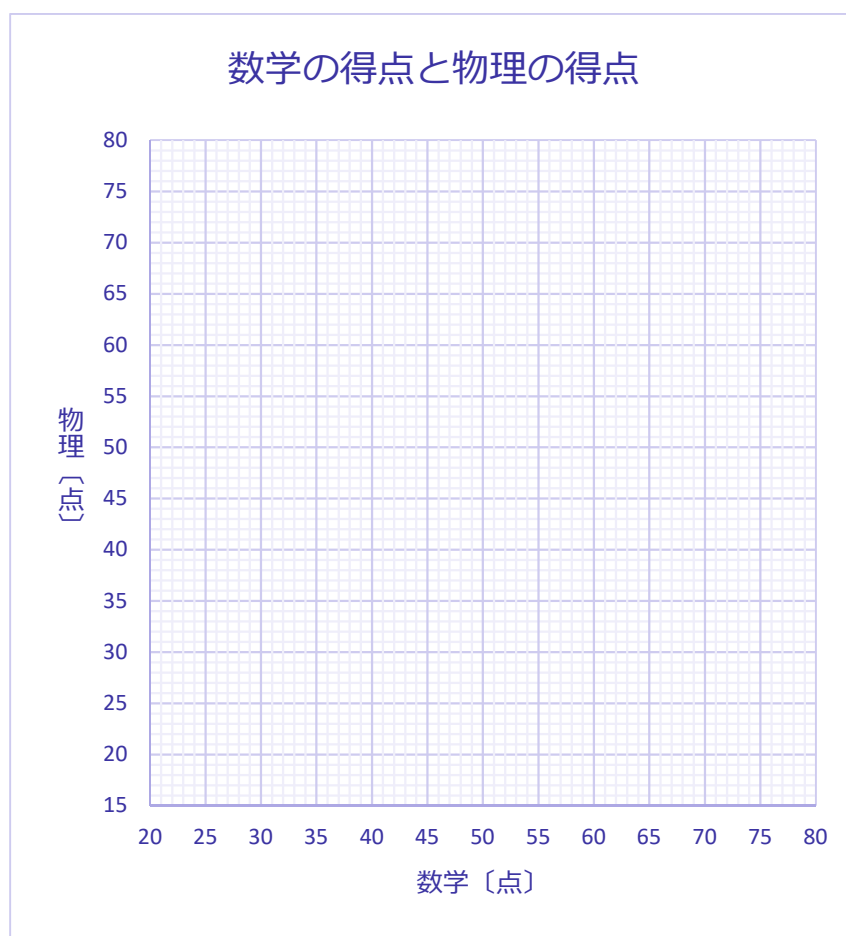
学生	1	2	3	4	5	6	7	8	9	10
数学	25	70	25	40	30	55	75	45	70	65
物理	25	55	30	55	20	45	75	45	75	75

①数学の得点、物理の得点のそれぞれの平均点を求めましょう。

数学の得点の平均点

物理の得点の平均点

②散布図を作成しましょう。





2 相関関係について

2変量データにおいて、一方が増加すると他方が増加する場合には**正の相関関係**があるといい、一方が増加すると他方が減少する場合には**負の相関関係**があるといいます。それらの傾向が認められない場合には、相関関係がないといいます。

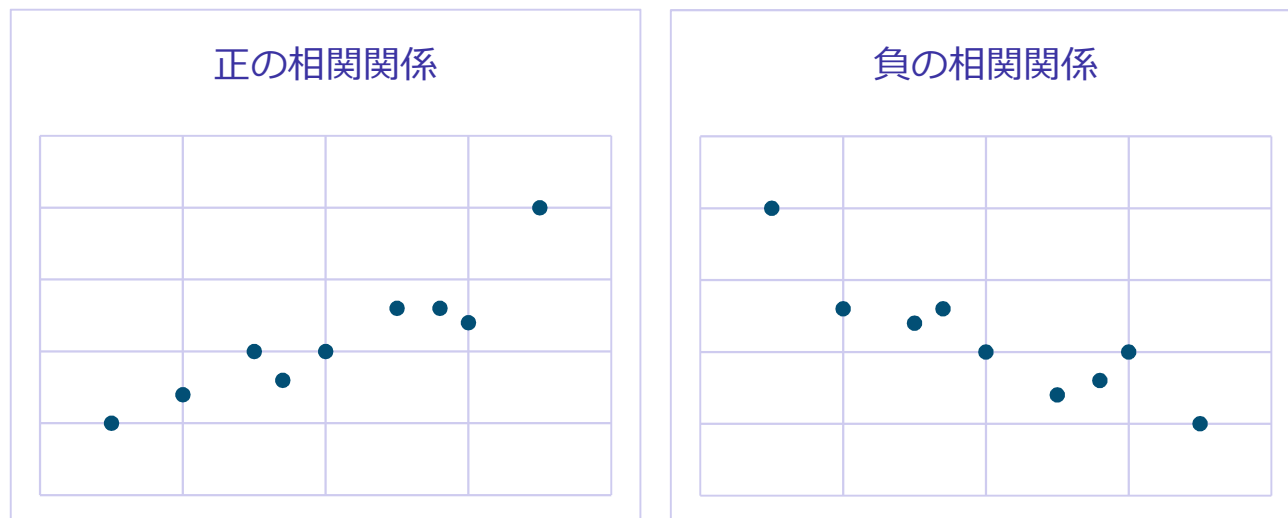


図4 正の相関関係と負の相関関係

(1) 正の相関関係

自分の経験から考えて、正の相関関係があると思うものを、いくつか挙げてみましょう。

(2) 負の相関関係

自分の経験から考えて、負の相関関係があると思うものを、いくつか挙げてみましょう。

3 相関関係と因果関係の違い

富士未来学で課題研究を進めていく過程で、自分が設定した仮説を検証する際に、相関関係を調べることもあります。しかし、相関があるからといってそれが原因と結果の関係にあるとは限らないことに注意が必要です。先ほどの例である数学の得点と物理の得点には正の相関関係がありますが、物理ができる原因が、数学ができることとは限りません。他にも原因があるかもしれません。



4 共分散について

相関関係を散布図からではなく、数値で表現する方法を考えていきます。ここでは、先ほどの散布図を利用して、**共分散**について学習します。共分散とは、二つのデータ間の関係を表す数値です。

先ほどの数学の得点と物理の得点の散布図を見てみましょう。平均値を基準に範囲を分けます。今回のデータでは、**正の相関関係**があるので、**右上と左下に集中**しています。

値から平均値を引いた値のことを**偏差**といいます。

偏差と偏差の符号の関係から点の位置がわかります。この性質を利用して、相関関係を数値で判断できるようになります。

まずは偏差を計算し、偏差の符号の関係を見ていきましょう。

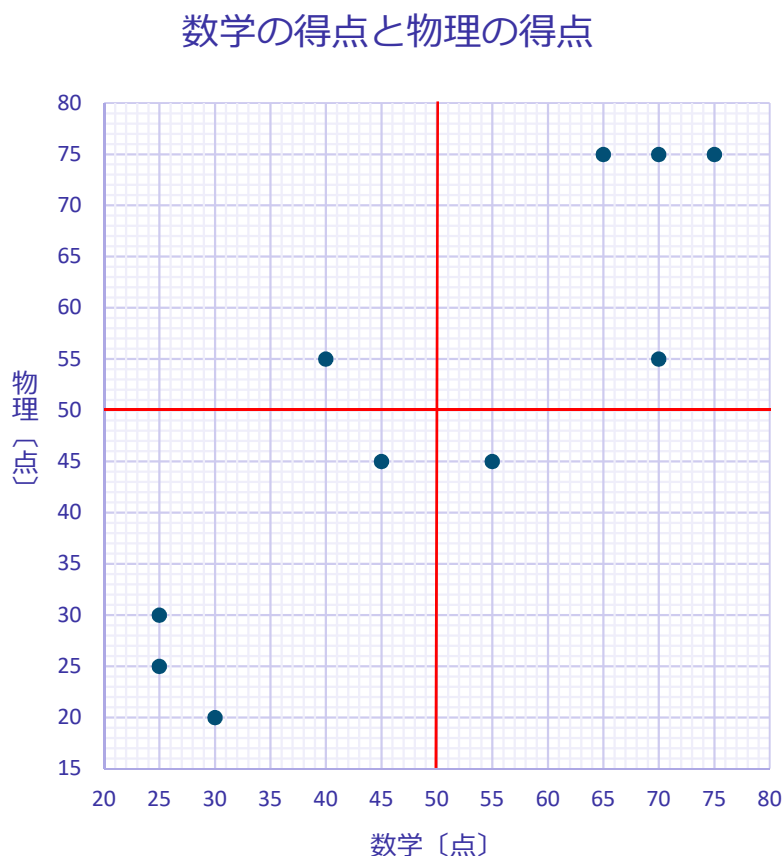


図5 数学の得点と物理の得点

①偏差を計算することで、下の表2を完成させましょう。

表2 偏差の符号の関係

学生	1	2	3	4	5	6	7	8	9	10
数学	25	70	25	40	30	55	75	45	70	65
偏差	-25	20	-25	-10						
物理	25	55	30	55	20	45	75	45	75	75
偏差	-25	5	-20	5						
点の位置	左下	右上	左下	左上						
偏差の符号の関係 同符号か異符号か	同	同	同	異						



②点の位置と偏差の符号の関係で分かることを書きましょう。

③正の相関関係があることと偏差の符号の関係について分かることを書きましょう。

共分散は、「偏差同士の積の平均」で定義されます。

④正の相関関係があるとき、計算した共分散の符号はどのようなになるでしょうか。

理由とともに書きましょう。

5 ルーブリックによる自己評価

育成したい資質・能力	グラントデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
挑戦力	試行錯誤	主体的に学習に取り組む態度	1と2の記述	1と2において、1で散布図を作成したり、2で相関関係について記述したりしている。	1と2において、1で散布図を作成していない、または、2で相関関係について記述していない。	1と2において、1で散布図を作成していない、かつ、2で相関関係について記述していない。		
理数的解決力	分析解析	思考・判断・表現	4の記述	4において、正の相関関係と共分散の符号の関係を、①から③で順を追って理解し、④で理由とともに記述している。	4において、①から点の位置と偏差の符号の関係を見いだして②を記述しているが、③を記述していない。	4において、①で偏差の計算や符号の判断ができず、表を埋めていない。		

引用文献

- (1) 東京都立富士高等学校・東京都立富士高等学校附属中学校（2021）『令和3年度スーパーサイエンスハイスクール研究開発実施計画書【開発型・実践型】』



データ分析講座Ⅱ②

ループブックを見返しながら、自己の到達度を確認し、改善を図りましょう。

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
挑戦力	試行錯誤	主体的に学習に取り組む態度	8 (1) の記述	8 (1) において、分散の大きさと平均を基準とした、ばらつき具合の関係を偏差の大きさの割合に着目して、説明している。	8 (1) において、分散の大きさと平均を基準とした、ばらつき具合の関係を説明しているが、偏差の大きさの割合に着目していない。	8 (1) において、分散の大きさと平均を基準とした、ばらつき具合の関係を説明していない。		
理数的解決力	分析解析	思考・判断・表現	1 1 (2) の記述	1 1 (2) において、そこまでに求めた共分散や分散の値を利用して、相関係数を、正しい値になるまで計算している。	1 1 (2) において、そこまでに求めた共分散や分散の値を利用して、相関係数を、正しい値になるまで計算していない。	1 1 (2) において、共分散や分散が求められておらず、相関係数の計算をしていない。		

6 共分散についての振り返り

共分散は、「偏差同士の積の平均」で定義されます。

偏差とは、値から平均値を引いた値のことです。右の散布図のデータについて、偏差同士の積の符号を確認しましょう。また、共分散の定義を参考に、共分散の符号を考えましょう。

①右上のデータ

②左上のデータ

③左下のデータ

④右下のデータ

⑤共分散の符号を予想しましょう。

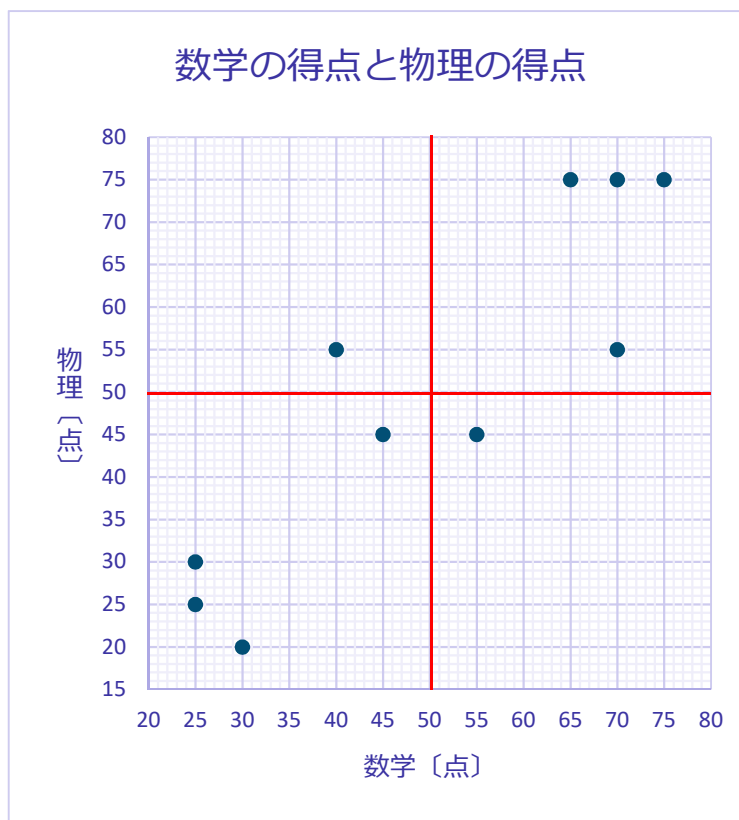


図 6 数学と物理の得点



7 共分散の定義

(1) 共分散の定義

表3 変量 x と変量 y

	1	2	3	...	n
x	x_1	x_2	x_3		x_n
y	y_1	y_2	y_3		y_n

変量 x と変量 y の**共分散** s_{xy} の定義を式で表すと、次のようになります。

$$s_{xy} = \frac{1}{n} \{ (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y}) \}$$

ここで、 $\bar{x}$ と $\bar{y}$ はそれぞれ平均を表しているとします。

(2) 共分散の計算

表4 数学と探究の得点と偏差

学生	1	2	3	4	5	6	7	8	9	10
数学	40	75	65	75	20	60	50	70	30	55
偏差										
探究	40	70	60	75	20	65	55	75	30	50
偏差										

表4のデータを用いて、共分散を計算しましょう。

数学の得点の平均点

探究の得点の平均点



8 分散についての振り返り

(1) 分散の定義

分散は、1変量について、各データの平均を基準としたばらつき具合を表す値です。偏差の2乗の平均を分散といいます。**分散 s^2** を式で表すと次のようになります。

$$s^2 = \frac{1}{n} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\}$$

※偏差の平均は必ず0になるので、偏差の平均を基準としたばらつき具合を表すことができません。

分散の値が大きいほど、平均を基準としたばらつき具合が大きいといえます。その理由を

分散の式を利用して説明しましょう。

(2) 分散の計算

表4のデータを用いて、数学の得点の分散を計算しましょう。

表4のデータを用いて、探究の得点の分散を計算しましょう。



9 標準偏差についての振り返り

(1) 標準偏差の定義

標準偏差は、1変量について、各データの平均を基準としたばらつき具合を表す値です。偏差の2乗の平均を分散といいます。その分散の値の正の平方根が標準偏差です。標準偏差 s_x を式で表すと次のようになります。

$$s_x = \sqrt{\frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\}}$$

※2乗(平方)すると a になる数を、 a の平方根といいます。例えば、2は2乗すると4になり、 -2 も2乗すると4になります。つまり、4の平方根は2と -2 の二つです。また、2の正の平方根は、記号 $\sqrt{\quad}$ を用いて、 $\sqrt{2}$ と表します。読み方は「ルート2」で、実際の値は、 $\sqrt{2} = 1.41421356\cdots$ で、循環しない無限小数(無理数)であることが知られています。

分散と標準偏差の単位はなんでしょうか。分散は計算過程で2乗されるので、データの単位が2乗されると考えられます。例えば、データの単位が〔点〕であるとき、分散の単位は〔点²〕となります。そこで、平方根を考えることによって、標準偏差の単位は〔点〕となり、データの単位と同じものになります。分散では単位を省略しますが、標準偏差では単位を明記します。共分散の単位も分散と同様に考えます。

(2) 標準偏差の計算

8 (2) の結果から、数学の得点の標準偏差を求めましょう。

8 (2) の結果から、探究の得点の標準偏差を求めましょう。

10 単位について

分散と標準偏差は、どちらも平均を基準としたばらつき具合を表しています。単位がデータと同じものになる標準偏差が、より分かりやすい指標になります。共分散についても、数値が大きくなる場合があります。分散、標準偏差、共分散は単位の影響を受けます。分散や標準偏差は、単位が a 倍になるとそれぞれ a^2 倍、 a 倍となります。共分散は、単位が a 倍、 b 倍になると ab 倍になります。

単位の影響を受けないものさしとして、次に扱う相関係数を考えます。



1.1 相関係数について

(1) 相関係数の定義

相関係数は、2変量について、各データの相関関係を数値で表した値です。相関係数は共分散をそれぞれの標準偏差で割ることで得られます。**相関係数 r** を式で表すと、次のようになります。**相関係数に単位はありません。**

$$r = \frac{S_{xy}}{S_x S_y}$$

$$\text{相関係数} = \frac{(x \text{ と } y \text{ の共分散})}{(x \text{ の標準偏差}) \times (y \text{ の標準偏差})}$$

式から、相関係数の符号は共分散の符号と一致することがわかります。正の相関関係があるときは、相関係数の符号は正になり、負の相関関係があるときは、相関係数の符号は負になります。相関係数 r には、次の性質があります。

- ① 相関係数 r は、 $-1 \leq r \leq 1$ の範囲にあります。
- ② r の値が 1 に近いとき、正の相関関係があるといえます。
- ③ r の値が -1 に近いとき、負の相関関係があるといえます。
- ④ r の値が 0 に近いとき、(直線的な) 相関関係がないといえます。

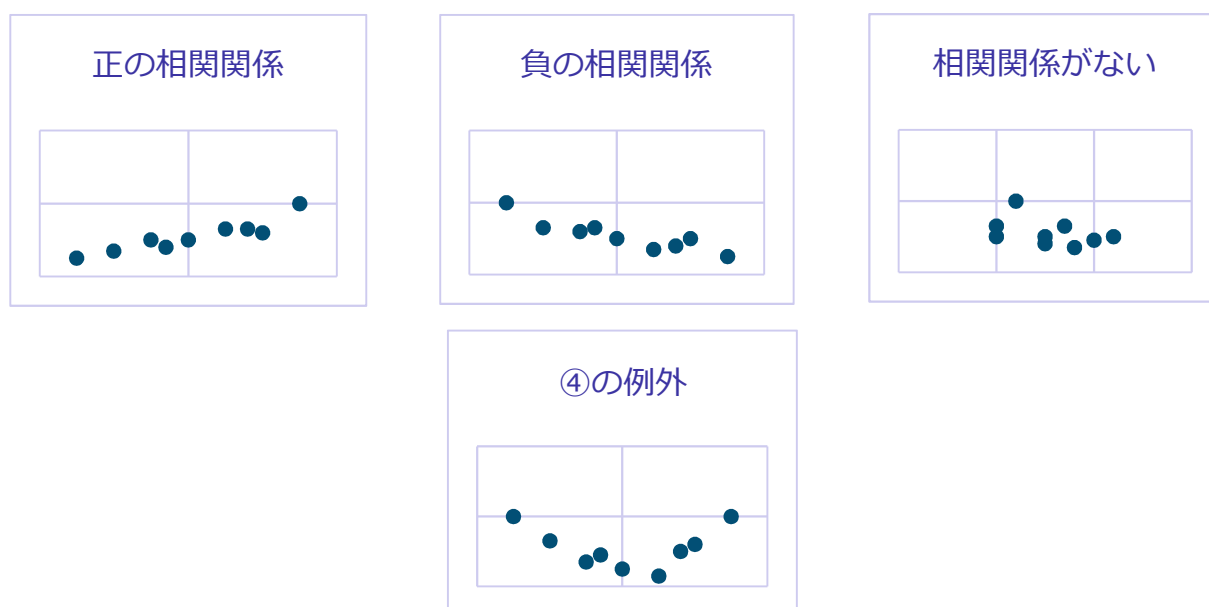


図7 いろいろな相関関係

相関係数は、直線的な相関関係を判断できます。図7の下の方のように、相関係数が0に近くても、相関関係がないといえない場合もあります。注意が必要です。



(2) 相関係数の計算

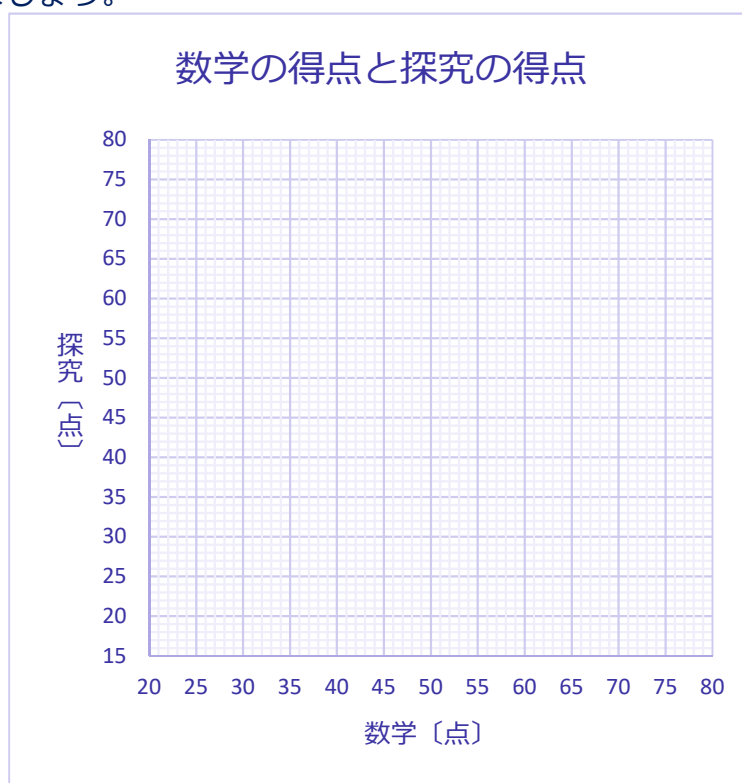
9ページの表4における数学と探究の得点の相関係数を計算しましょう。小数第4位を四捨五入しましょう。

(3) 相関係数と散布図

表5 数学と探究の得点

学生	1	2	3	4	5	6	7	8	9	10
数学	40	75	65	75	20	60	50	70	30	55
探究	40	70	60	75	20	65	55	75	30	50

散布図を作成しましょう。





1 2 相関関係の注意点

前回の講座でも触れましたが、相関関係があるからといって、因果関係があるとはいえません。今回扱った数学の得点と探究の得点にでも、数学ができることが、探究ができることの原因かどうかは分かりませんし、その逆も同様です。

例えば、ある日の冷たい飲み物の売り上げと熱中症になる人の数を散布図で表すと、正の相関関係があることが分かりました。では、冷たい飲み物を売れば売ほど、熱中症になる人が増えるという因果関係があるのでしょうか。この日は気温が上昇していたとすると、熱中症になる原因は、気温の上昇であると考えの方が自然ではないでしょうか。

上の例のように、相関関係があることが、因果関係があることを保証しませんし、他にも原因がある可能性もあります。ただし、相関関係を調べることの意味がないわけではありません。因果関係があるためには、相関関係がある必要があります。相関関係があるものの中から因果関係を探っていくのです。

1 3 ルーブリックによる自己評価

育成したい資質・能力	グラントデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
挑戦力	試行錯誤	主体的に学習に取り組む態度	8 (1) の記述	8 (1) において、分散の大きさと平均を基準としたばらつき具合の関係を、偏差の大きさの割合に着目して、説明している。	8 (1) において、分散の大きさと平均を基準としたばらつき具合の関係を、説明しているが、偏差の大きさの割合に着目していない。	8 (1) において、分散の大きさと平均を基準としたばらつき具合の関係を説明していない。		
理数的解決力	分析解析	思考・判断・表現	1 1 (2) の記述	1 1 (2) において、そこまでに求めた共分散や分散の値を利用して、相関係数を、正しい値になるまで計算している。	1 1 (2) において、そこまでに求めた共分散や分散の値を利用しているが、相関係数を、正しい値になるまで計算していない。	1 1 (2) において、共分散や分散が求められておらず、相関係数の計算をしていない。		

引用文献

- (1) 東京都立富士高等学校・東京都立富士高等学校附属中学校（2021）『令和3年度スーパーサイエンスハイスクール研究開発実施計画書【開発型・実践型】』



データ分析講座Ⅱ③

ルーブリックを見返しながら、自己の到達度を確認し、改善を図りましょう。

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
理数的解決力	分析解析	思考・判断・表現	1 7 ③④の記述	1 7の③と④において、自力または解説を参考にして、標準偏差や共分散を求めている。	1 7の③と④において、自力または解説を参考にして、標準偏差や共分散のどちらか一方を求めている。	1 7の③と④において、標準偏差や共分散を求めている。		
理数的解決力	分析解析	思考・判断・表現	1 7 ⑤⑦の記述	1 7の⑤と⑦において、自力または解説を参考にして、回帰直線を求め、荷重 $x = 5$ に対する伸び y を求めている。	1 7の⑤と⑦において、自力または解説を参考にして、回帰直線を求めているが、荷重 $x = 5$ に対する伸び y を求めている。	1 7の⑤において、回帰直線を求めている。		

1 4 回帰分析とは

変数 x と変数 y の間に、相関だけではなく、 y が x で説明できるような関係があるとき、 x を独立変数（説明変数）、 y を従属変数（目的変数）といいます。例えば、重りの重さ x 〔g〕とばねの伸び y 〔cm〕、年齢 x 〔歳〕と血圧 y 〔mmHg〕などがあります。

図1のように、変数 X と変数 Y には正の相関関係があり、およそ1次式（ $y = ax + b$ ）の関係があることが読み取れます。この直線を回帰直線といいます。データに回帰直線をあてはめて解釈することを、回帰分析といいます。特に、独立変数が一つであるとき、単回帰分析といいます。

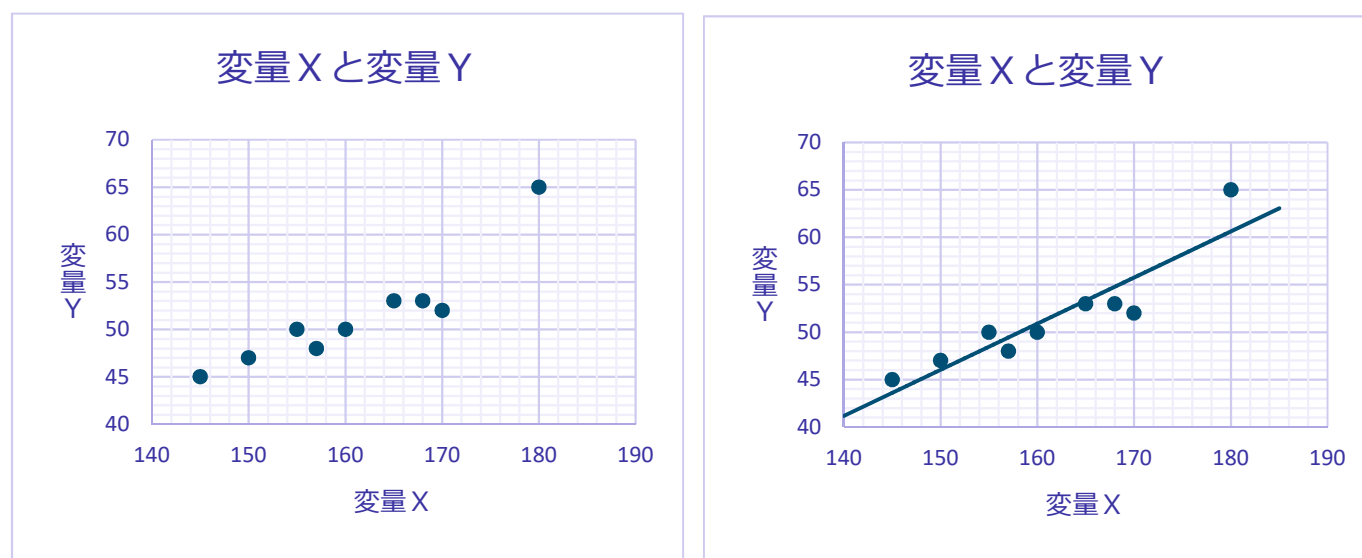


図8 1次式の関係



1.5 回帰と回帰直線

図8の右図のように、いくつかの点の配列を1本の曲線で代表することを**回帰**といいます。特に、1本の直線で回帰するとき、その直線 $y = ax + b$ を y の x への**回帰直線**といいます。また、傾きと切片を**回帰係数**といいます。

回帰直線は、それぞれの点の近くを通るようにしなければなりません。それぞれの点の近くを通るような回帰直線はどのように求めたらよいのでしょうか。

1.6 回帰直線の公式

変量 x と変量 y の平均をそれぞれ $\bar{x}$ と $\bar{y}$ 、標準偏差をそれぞれ s_x と s_y 、共分散を s_{xy} 、相関係数を r とします。

直線 $y = ax + b$ を y の x への回帰直線とすると、

$$a = \frac{s_{xy}}{s_x^2} = \frac{(x \text{ と } y \text{ の共分散})}{(x \text{ の分散})}, \quad b = \bar{y} - a\bar{x} \quad \times \quad a = r \frac{s_y}{s_x}$$

証明は、大学数学の内容を利用しますので、考え方の確認に留めます。図9を見てみましょう。各点の y 座標 y_i と直線上の値 $\hat{y}_i = ax_i + b$ の差（残差といいます）の2乗の値の和が最小になるように考えます。この方法を、**最小二乗法**といいます。

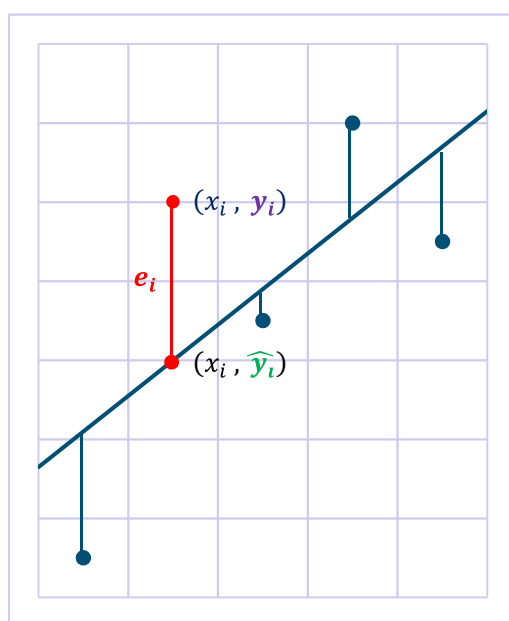


図9 最小二乗法

実現値（実際のデータ）である「 y_i 」と、予測値（直線上のデータ）である「 $\hat{y}_i$ 」の差が残差「 e_i 」です。最小二乗法では、残差の2乗の和が最小になるように考えます。

最小二乗法には偏微分などの計算が必要になります。この講座では、証明は省略し、結果を利用することにします。



1.7 回帰直線を求める

次の表6は、ばねの変位を測定して得られたデータである。

表6 ばねの変位

	1	2	3	4	5
荷重 x	0	1	2	3	4
伸び y	10	14	21	26	29

①荷重、伸びのそれぞれの平均値を求めましょう。

荷重の平均値 $\bar{x}$

伸びの平均値 $\bar{y}$

②荷重の分散 s^2 を求めましょう。

$$s^2 = \frac{1}{n} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\}$$

③荷重の標準偏差 s_x を求めましょう。

$$s_x = \sqrt{\frac{1}{n} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\}}$$

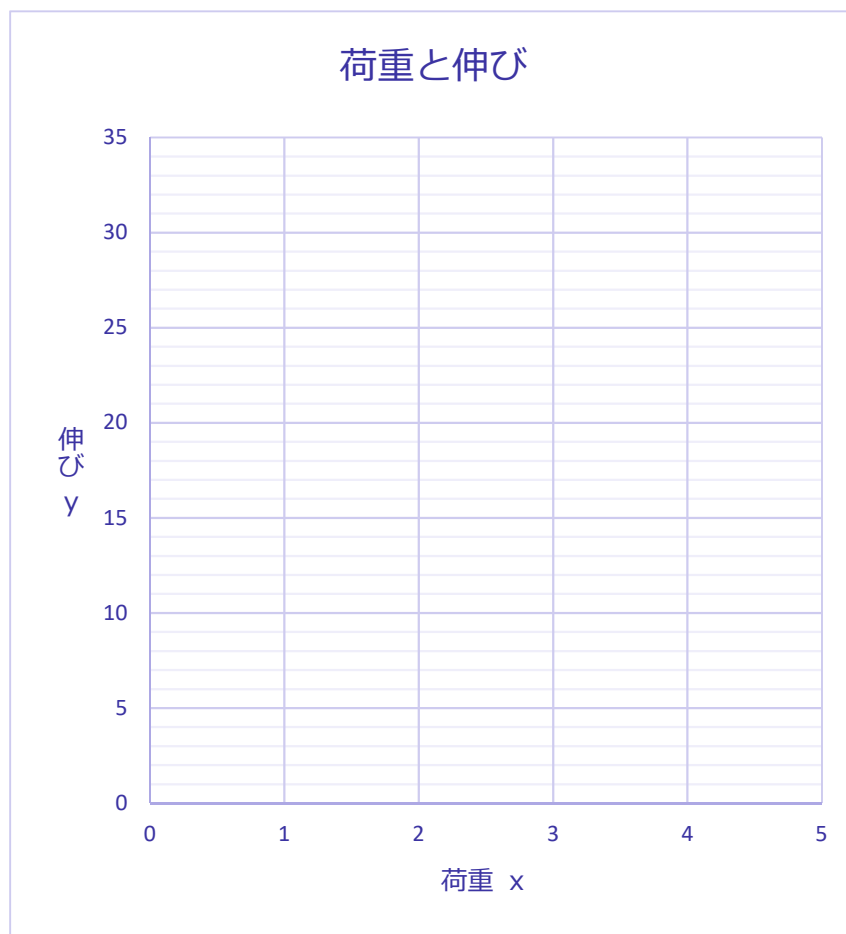
④荷重と伸びの共分散 s_{xy} を求めましょう。

$$s_{xy} = \frac{1}{n} \{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})\}$$



⑤回帰直線の傾き a と切片 b を求め、回帰直線の方程式を求めましょう。

⑥散布図を作成し、回帰直線のグラフもかきましょう。



⑦求めた回帰直線の方程式を用いて、荷重 $x = 5$ に対する伸び y を求めましょう。



1 8 1 変量から 2 変量、多変量へ

データ分析講座Ⅰでは、平均値から始まり標準偏差までの 1 変量について、データ分析講座Ⅱでは、共分散から始まり相関関係や直線回帰までの 2 変量の関係を扱いました。これまでに学んだことを活用して、データ分析講座Ⅲでは、多変量解析の基礎に触れます。

富士未来学の課題研究では、相関関係から因果関係を考えたり、実験のデータから回帰直線を利用して予測したりしながら、自ら設定した仮説を検証していくことになります。さらには、検定などの推測統計を活用して、得られた数値を根拠に結論を導きます。検定についてもデータ分析講座Ⅲで扱う予定です。

1 9 視覚的な説明

共分散（相関係数）、分散、直線回帰の視覚的な説明をします。あくまでも補足説明です。

（1）共分散（相関係数）の符号の視覚的な説明（青い点はデータを表しています）

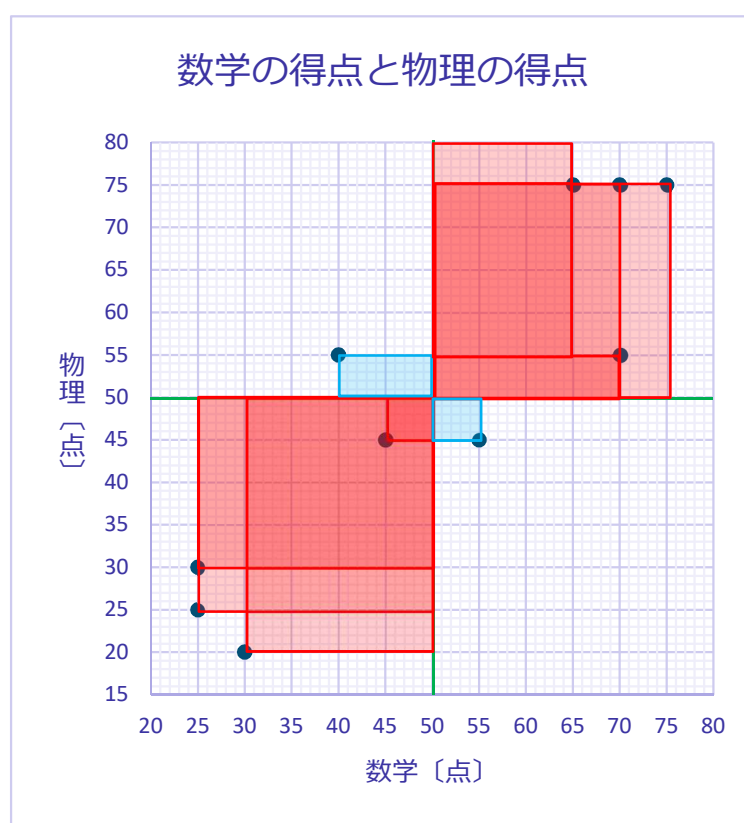


図 1 0 共分散の符号の視覚化

緑の直線はそれぞれの平均値を表しています。偏差の積は、縦×横となるので、四角形の面積を求めていると考えることができます。赤い四角形の面積の和から青い四角形の面積の和を引いた値の符号が、共分散（相関係数）の符号になります。



(2) 分散と直線回帰の視覚的な説明（青い点はデータを表しています）

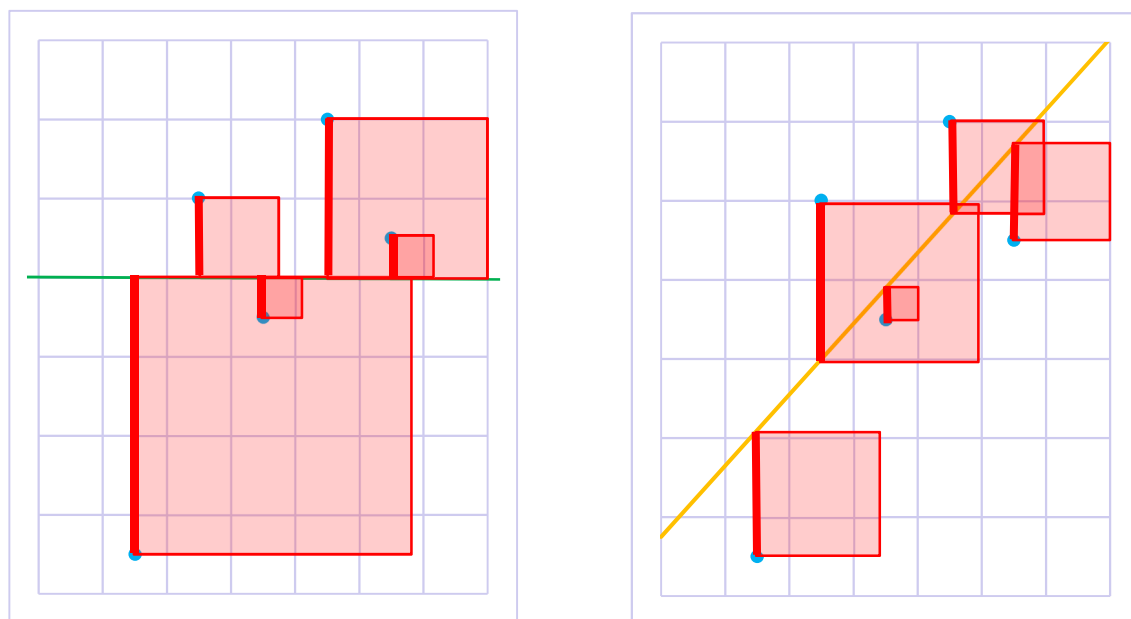


図 1.1 分散や直線回帰を面積で表現

左の図で、緑の直線は平均値を表しています。分散の大きさは、赤い正方形の面積の和として捉えることができます（正確にはその平均値）。平均値からの散らばりが小さいほど、それぞれの赤い正方形が小さくなり、面積の和も小さくなることが分かります。

右の図で、橙色の直線は回帰直線を表しています。赤い正方形の面積の和が小さくなるように回帰直線を引いています。残差の2乗が赤い正方形の面積を表しています。

2.0 相関係数の信頼性

相関係数は、データの数によって、信頼性が変わります。例えば、全データから取り出した7組のデータで求めた相関係数が0.85であったとします。この相関係数は、真の相関係数（全データで求めた相関係数）とどのような関係があるのでしょうか。真の相関係数は0.85といえるのでしょうか。

7組のデータで求めた相関係数が0.85であったとき、「100回中95回は、区間0.26～0.97が真の相関係数を含む」ことしかわかりません。この区間を95%信頼区間といいますが、この精度では、十分信頼できるといえるのでしょうか。

データの数を増やすと、信頼区間の幅が狭くなり、精度があがっていくことが知られています。つまり、データの数、信頼性に大きく影響を与えるのです。

この話を理解するためには、推定の知識が必要になります。データ分析講座Ⅲや統計分析講座で学習します。



2 1 ルーブリックによる自己評価

育成したい資質・能力	ルーブリックの観点	評価の観点	評価の対象	高度に達成されている	達成されている	一部に課題あり	自己評価	教員による評価
				A	B	C		
理数的解決力	分析解析	思考・判断・表現	1 7 ③④の記述	1 7の③と④において、自力または解説を参考にして、標準偏差や共分散を求めている。	1 7の③と④において、自力または解説を参考にして、標準偏差や共分散のどちらか一方を求めている。	1 7の③と④において、標準偏差や共分散を求めている。		
理数的解決力	分析解析	思考・判断・表現	1 7 ⑤⑦の記述	1 7の⑤と⑦において、自力または解説を参考にして、回帰直線を求め、荷重 $x = 5$ に対する伸び y を求めている。	1 7の⑤と⑦において、自力または解説を参考にして、回帰直線を求めているが、荷重 $x = 5$ に対する伸び y を求めている。	1 7の⑤において、回帰直線を求めている。		

引用文献

- (1) 大村平 (2018) 「相関係数はこれだ」「直線で回帰する」『多変量解析のはなし【改訂版】—複雑さから本質を探る—』, 日科技連, pp.43-45, pp.75-79.
- (2) 高遠節夫ほか (2019) 「データの整理」『新確率統計』, 大日本図書, pp.48-53.
- (3) 東京都立富士高等学校・東京都立富士高等学校附属中学校 (2021) 『令和 3 年度スーパーサイエンスハイスクール研究開発実施計画書【開発型・実践型】』