第２章－６　異なるデータの関係性を客観的に把握したい

例えば　新潟県の人口動態について調べたところ、婚姻数と出生数の間には一方が増えると他方も増えるという関係があることがみてとれた。

(表1)新潟県の出生数と婚姻数の推移

	H23	H24	H25	H26	H27	H28	H29	H30	R1	R2	R3	R4	R5	R6
婚姻数(件)	10,278	10,219	9,965	9,955	9,437	9,312	8,916	8,612	8,742	7,570	7,088	6,823	6,262	6,352
出生数(人)	17,667	17,476	17,066	16,480	16,340	15,737	14,967	14,509	13,640	12,981	12,608	11,732	10,916	9,941

資料:新潟県福祉保健部「令和6年人口動態統計(確定数)の概況(新潟県版)」

例①両者について客観的に関係性を検証できないか？

例②両者に関係性があると言えるとき、婚姻数が 12,000 件に増加したときの出生数を予測したい。

分析の仕方

○相関分析（例①の分析）

相関：２つの変数の間に（直線的な）関係がある（※）かどうかのこと。

※一方が増えると他方も増える、一方が減ると他方も減るといった関係

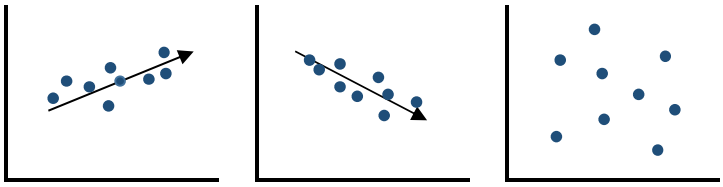
（例）身長と体重、所得と消費額

関係の強さを「相関係数」で検証することができる。

手順１　散布図の作成：<u>散布図では、視覚的に２つのデータの関係性をみることができる</u>

●データの見方（散布図）

データの分布状況により、おおむね以下のようなことがいえる。



「正の相関あり」　「負の相関あり」　「相関なし(無相関)」

手順２　相関係数の検証：<u>相関係数では、データ同士の関係性の強さを数値でみることができる</u>

●データの見方（相関係数）

絶対値が1に近いほど、数値の関係がある(相関がある)と解釈できる。

(相関の強さの目安の絶対的な基準はないが、参考として下記の表のような見方がある)

相関係数の絶対値	相関の強さの目安
0.7～	強い相関
0.4～0.7	中程度の相関
0.2～0.4	弱い相関
～0.2	ほとんど無相関

※相関係数がプラスであれば正の相関、マイナスであれば負の相関となる。

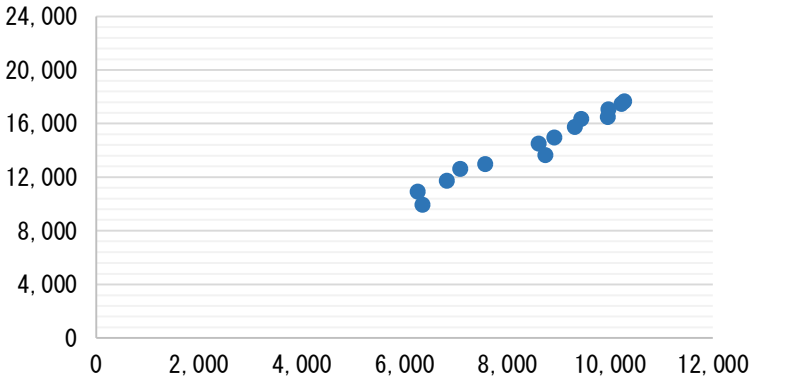
分析結果からわかること

分析結果１

表１のデータを使用して散布図を作成すると以下のようになる。

データの分布状況から、婚姻数と出生数の間には正の相関があるとみられる。

（図１）散布図　（新潟県の婚姻数と出生数の関係）



分析結果２

表１のデータから、エクセルの分析ツールを使用して相関係数を求める（※）と、以下のようになる。（※第２章－８「エクセル分析ツールの使い方」参照）

相関係数は約 0.98 となり、0.7 を超えるため、強い正の相関があると考えられる。

(表2)エクセルの分析ツールによる相関分析結果

	婚姻数(件)	出生数(人)
婚姻数(件)	1	
出生数(人)	0.98186834	1

相関係数

⇒分析結果１、２より、婚姻数と出生数の間には強い正の相関があることがわかる。

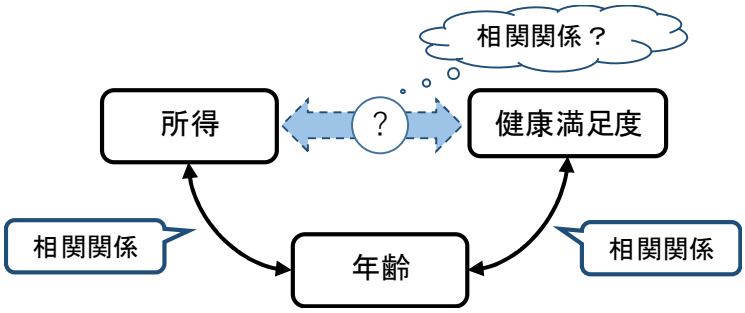
※相関をみる際の注意点

〇擬似相関に注意

・・・擬似相関とは、データ同士に直接の相関性がないのに、別の要因の影響などにより、見かけ上は相関関係があるようにみえる（相関係数の絶対値が高く出る）こと。

（例）所得と健康満足度に負の相関係数が大きく出た。所得が高い人ほど健康状態が悪い？

⇒この関係の間には、他の要因（年齢）の影響がある可能性がある。



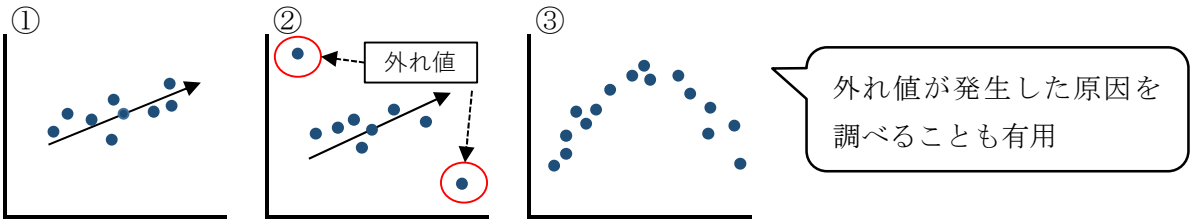
・・・ 一般的には、年齢が上がる→所得が上がる

年齢が上がる→健康満足度が下がる　という傾向があるとみられる

→「所得」と「年齢」、「健康満足度」と「年齢」の間に相関関係がみられた場合、「所得」と「健康満足度」の間の関係は擬似相関の可能性がある

⇒所得と健康満足度の間の相関関係を調べるためには、他の要因（年齢）を除去して考える必要がある。（△歳のときの所得と健康満足度の関係、とするなど）

〇相関係数は外れ値に影響される。また、直線的な関係以外は関係性を数値化できない。



※例①～③の相関係数の値は仮の値

・・・例②、③のような場合、相関係数のみをみると、ほとんど無相関と考えられる。

しかし、散布図を作成すると何らかの関係性がありそうにみえる。

⇒このため、分析手順１のように、まず散布図を作り、２つのデータの関係を確認することも重要となる。

分析の仕方

○回帰分析（例②の分析）

例②：県の出生数と婚姻数に関係性があることが分かった。統計データから、婚姻数が 12,000 件に増加したときの出生数を予測したい。

(表1)新潟県の出生数と婚姻数の推移

	H23	H24	H25	H26	H27	H28	H29	H30	R1	R2	R3	R4	R5	R6
婚姻数(件)	10,278	10,219	9,965	9,955	9,437	9,312	8,916	8,612	8,742	7,570	7,088	6,823	6,262	6,352
出生数(人)	17,667	17,476	17,066	16,480	16,340	15,737	14,967	14,509	13,640	12,981	12,608	11,732	10,916	9,941

資料:新潟県福祉保健部「令和6年人口動態統計(確定数)の概況(新潟県版)」

回帰分析（単回帰分析）：２つのデータの間に相関関係（直線的な関係）がある場合、その関係を直線の式（単回帰式：ｙ＝ａｘ＋ｂ）にあてはめる。この回帰式の一方の変数に値を代入することで、他方の変数の値の予測が可能となる。

手順１　回帰式を求める：データの関係性を式で表す

散布図またはエクセルの分析ツールにより求めることが可能。（※第２章－８「エクセル分析ツールの使い方」参照）

●データの見方（単回帰分析）

求めた回帰式の精度については、あらかじめ相関分析を行うことのほか、相関係数を２乗した値（R-2 乗値）により確認できる。

R-2 乗値が１に近いほど回帰式の精度がいいといえる。

⇒一般的には　$R^2 \geqq 0.5$　が目安とされることが多い。

※散布図では「R^2」、データ分析ツールでは「重決定 R2」と記載されている

※R-2 乗値の目安となる 0.5 というのは、相関係数の強い相関があるとされる目安の 0.7 を２乗した値（0.7×0.7＝0.49≒0.5）からきている。

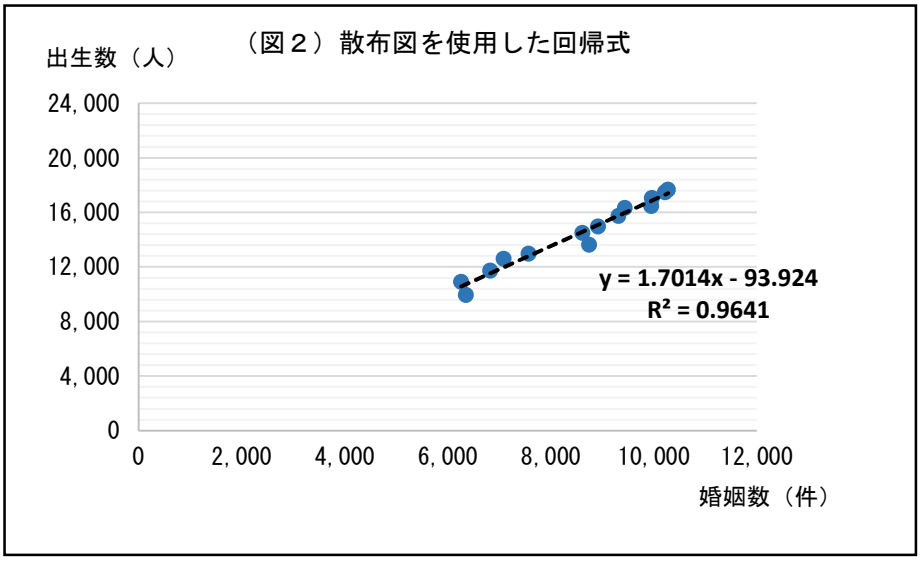
手順２　回帰式を使って値の予測をする：回帰式に数値を代入する

手順１で求めた回帰式に、設定した数値（今回は婚姻数）を代入することで、求めたい数値（今回は出生数）を計算することができる。

分析結果からわかること

分析結果１

表１のデータを使用して、回帰式を求めると、以下のようになる
（※第２章－８「エクセル分析ツールの使い方」４　回帰式の求め方　参照））



⇒表１のデータによる婚姻数（ｘ）と出生数（ｙ）の間の関係式は
ｙ＝1.7014ｘ－93.924　であることがわかる。
また、R-2 乗値は 0.9641 であり、上記回帰式の使用が可能と考えられる。

分析結果２

表１のデータを使用して求めた回帰式に、設定した婚姻数（ｘ＝12,000）を代入すると、
ｙ＝1.7014×12000－93.924＝20322.876
ｙ≒20,323 件　と予測できる。

⇒表１のデータから、婚姻数が 12,000 件に増加したときの出生数は 20,323 件と予測できる。

出生数（ｙ）を設定して、必要な婚姻数（ｘ）を求めるなど、代入する説明変数を入れ替えて予測することもできる。

参考

例のケースでは、２つのデータの関係を示す式を求めた（単回帰分析）が、あるデータ（目的変数）の数値に影響を与えていると思われるデータ（説明変数）が複数ある場合にも、それらの関係を式で表すことができる（＝**重回帰分析**）。

重回帰分析の回帰式は、

$y = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + b$

※重回帰分析もエクセルの分析ツールで行うことができる