データリテラシー

第10回　2変量データと相関

メディア情報コース
平居　悠（ひらい　ゆたか）

到達目標

- データサイエンス・AIが社会でどのように活用されているのかを理解する。
- データを適切に読み解き理解する力をつける。
- データを適切な可視化手法で他者に説明できるようになる。
- 表計算ソフトを使ってデータを集計・加工できるようになる。

前回

第１回　社会におけるデータ・AI利活用
第２回　表計算ソフトの基本操作
第３回　セルの参照と集計
第４回　度数分布表とヒストグラム
第５回　代表値と散らばりの尺度
第６回　データのグラフによる表現とデータ比較
第７回　データの分布と標本抽出
第８回　推定
第９回　分割表とクロス集計
第１０回　２変量データと相関
第１１回　時系列データ
第１２回　仮説検定
第１３回　回帰分析

前回の目標

多変量データについて分割表とクロス集計表を作成できるようになる

前回学んだこと

1. 分割表
2. クロス集計表
3. ピボットテーブル

多変量データ

ある集団に対して２種類以上の観測を行って得られた多変数のデータ

２変量データ（２次元データ）

2変数のデータ。2変量データを見やすくするには、分割表と散布図の二通りの方法がある。

学年	コース
4	観まち
4	観まち
2	政策
3	情報
3	経営
4	経営
4	政策
2	観まち
4	国際

分割表

2変量データの片方の変数のとりうる値を縦方向（列方向）の項目、もう片方の変数のとりうる値を横方向（行方向）の項目にして、縦と横の項目が交差（クロス）するマスに対応する度数を数えて入れた表

分割表

「合計」の欄は項目数に含めない。列や行の合計は**周辺度数**という。列の周辺度数の合計と行の周辺度数の合計は総数に一致する。

クロス集計表

分割表において度数ではなく、合計、平均、標準偏差などを集計した表。

ピボットテーブル

分割表やクロス集計表を作るための
表計算ソフトの機能。

前回の問い

- 分割表とは何か？
  - 2変量データを列と行に入れ、列と行の項目が交差するマスに対応する度数を数えて入れた表
- クロス集計表とは何か？
  - 合計、平均、標準偏差などを集計した表。
- 表計算ソフトで分割表やクロス集計表を作るには何を使うのか？
  - ピボットテーブルを用いる。

タイピング練習スケジュール

第１回　　ホームポジション
第２回　　ホームポジション
第３回　　ローマ字
第４回　　英語初級
第５回　　日本国憲法（trr試験）
第６回　　ホームポジション
第７回　　ローマ字
第８回　　英語初級
第９回　　日本国憲法（trr試験）
第１０回　ホームポジション（ボーナス練習）
第１１回　ローマ字
第１２回　英語初級
第１３回　日本国憲法（trr試験）

タイピング練習 (ホームポジション) 10分間

1. ブラウザを起動し、https://www.koeki-prj.org/trr/ に繋ぐ。
2. 学籍番号（Cは大文字、省略なし８桁）を入力する。
3. Koeki MAILに届いたパスコードをPasscode: 欄に入力する。

タイピング練習 (ローマ字) 10分間

1. ブラウザを起動し、https://www.koeki-prj.org/trr/ に繋ぐ。
2. 学籍番号（Cは大文字、省略なし８桁）を入力する。
3. Koeki MAILに届いたパスコードをPasscode: 欄に入力する。

タイピング練習 (英語初級) 10分間

1. ブラウザを起動し、https://www.koeki-prj.org/trr/ に繋ぐ。
2. 学籍番号（Cは大文字、省略なし８桁）を入力する。
3. Koeki MAILに届いたパスコードをPasscode: 欄に入力する。

タイピング練習 (日本国憲法) 10分間

1. ブラウザを起動し、https://www.koeki-prj.org/trr/ に繋ぐ。
2. 学籍番号（Cは大文字、省略なし８桁）を入力する。
3. Koeki MAILに届いたパスコードをPasscode: 欄に入力する。

ホームポジション


左手でタイプするキー
右手でタイプするキー
小指
薬指
中指
人差指
人差指
中指
薬指
小指
漢字
Back space
Tab
Enter
Caps Lock
Shift
Shift
Ctrl
Fn
Alt
Space
Alt
Fn
Ctrl
左手の人差指から小指までのホームポジション
両手の親指のホームポジション
右手の人差指から小指までのホームポジション

今回

第１回　社会におけるデータ・AI利活用
第２回　表計算ソフトの基本操作
第３回　セルの参照と集計
第４回　度数分布表とヒストグラム
第５回　代表値と散らばりの尺度
第６回　データのグラフによる表現とデータ比較
第７回　データの分布と標本抽出
第８回　推定
第９回　分割表とクロス集計
第１０回　２変量データと相関
第１１回　時系列データ
第１２回　仮説検定
第１３回　回帰分析

今回の目標

2変量データの相関について定量的に分析できるようになる。

今回学ぶこと

1. 相関関係
2. 相関係数
3. 因果関係と相関関係

今回の問い

- 「相関関係がある」とはどのような場合か？
- 相関関係を数量的に表す指標とは何か？
- 相関関係のある2つの物事について解釈する際に気を付けるべきことは何か？

今回学ぶこと

1. 相関関係
2. 相関係数
3. 因果関係と相関関係

相関関係

2変量X, Yのデータがあるとき、XとYの間の関係のことを**相関関係**という。

相関関係

統計学では、XとYの間に比例関係（直線関係）に近い傾向が見られることを「相関関係がある」という。

相関の正負

- **正の相関**：*X*が増加するにつれて*Y*は増加する。
- **負の相関**：*X*が増加するにつれて*Y*は減少する。

相関の程度

相関の程度は強い相関、弱い相関で表現する。

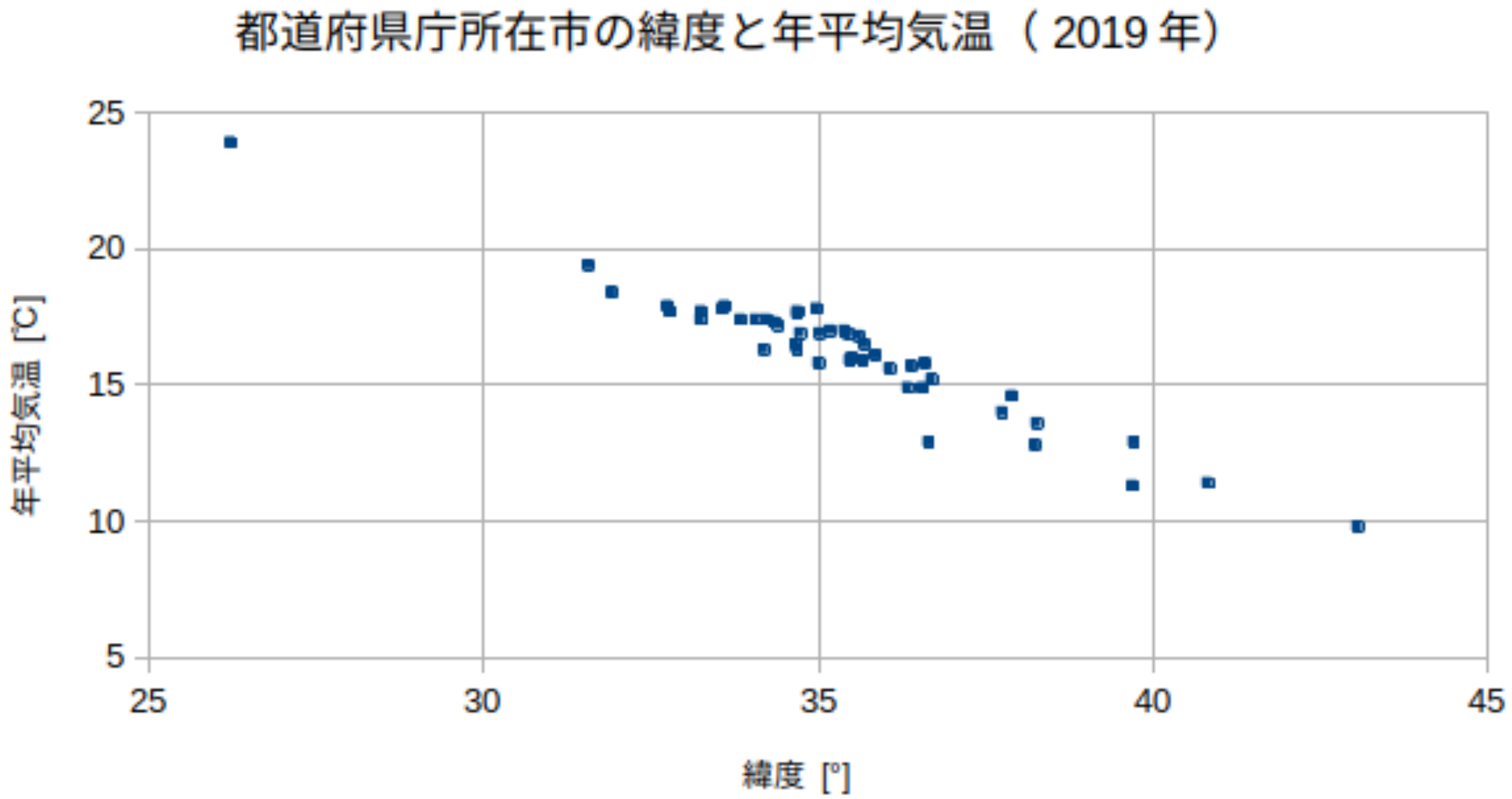
- **強い相関**：*X*と*Y*が直線的な関係に近い。
- **弱い相関**：*X*と*Y*が直線的な関係から離れている。

相関の程度

極端な場合は完全相関や無相関という。

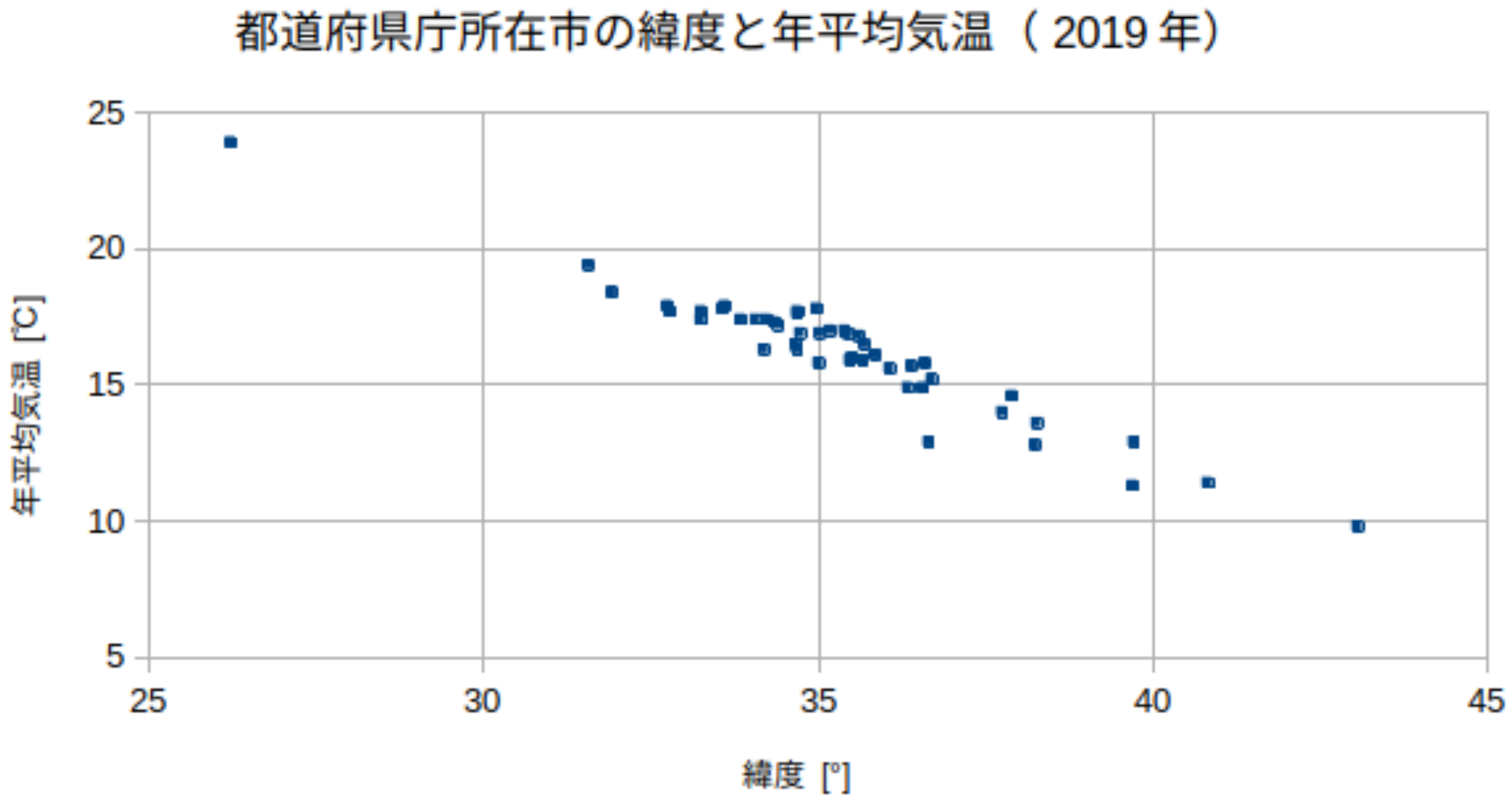
- **完全相関**：*X*と*Y*が完全に直線的な関係にある。
- **無相関**：*X*と*Y*がバラバラで全く直線的な関係にない。

相関の例



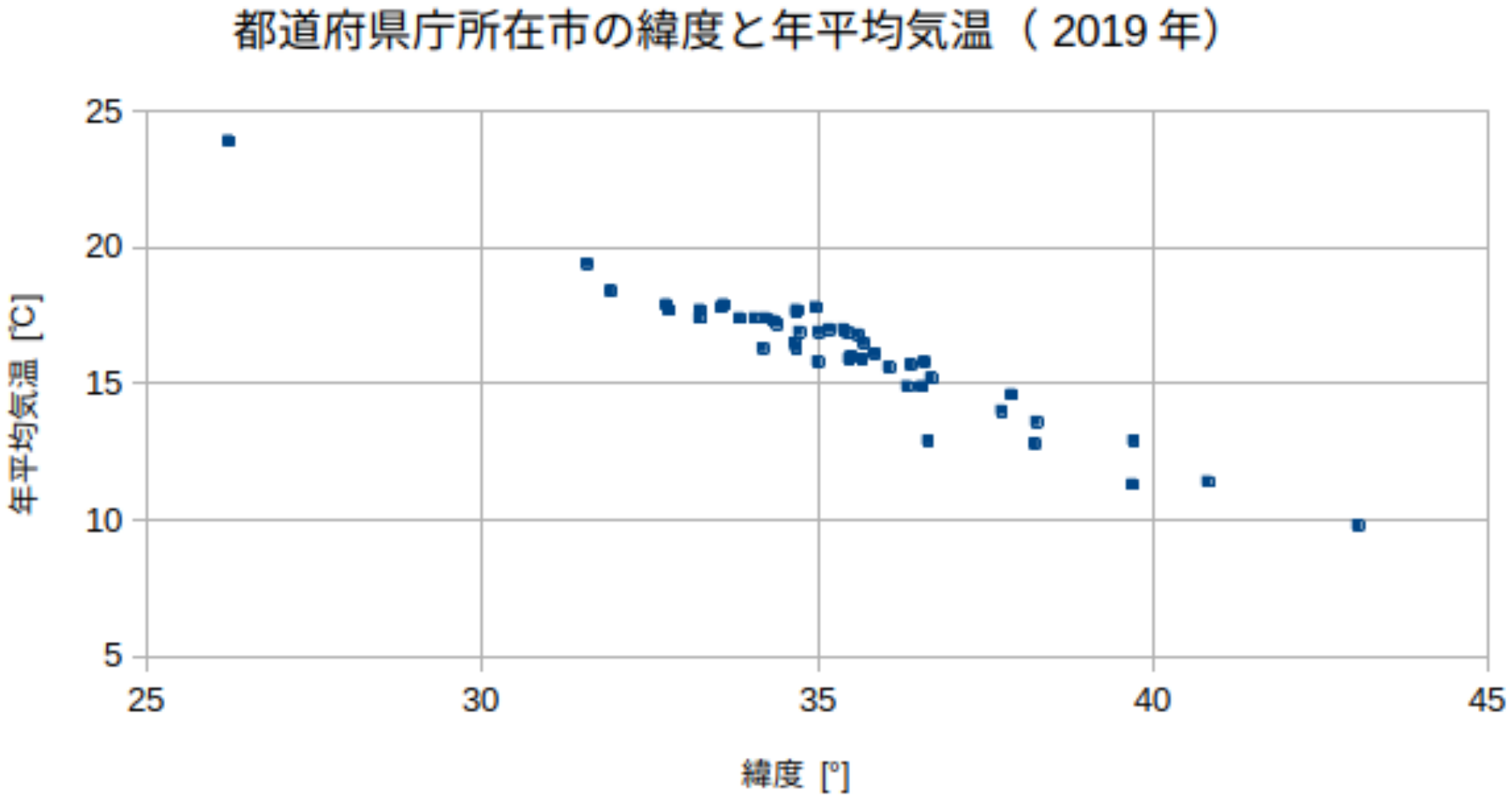
ここから何がわかるか？

相関の例



緯度が高くなるにつれて年平均気温は低くなる

相関の例



各点はほぼ直線上に並んでいるので、緯度と年平均気温は「強い負の相関」を持つといえる。

今回の問い

- **「相関関係がある」とはどのような場合か？**
  - **2変量が比例関係に近い傾向が見られる場合。**
- 相関関係を数量的に表す指標とは何か？
- 相関関係のある2つの物事について解釈する際に気を付けるべきことは何か？

今回学ぶこと

1. 相関関係
2. **相関係数**
3. 因果関係と相関関係

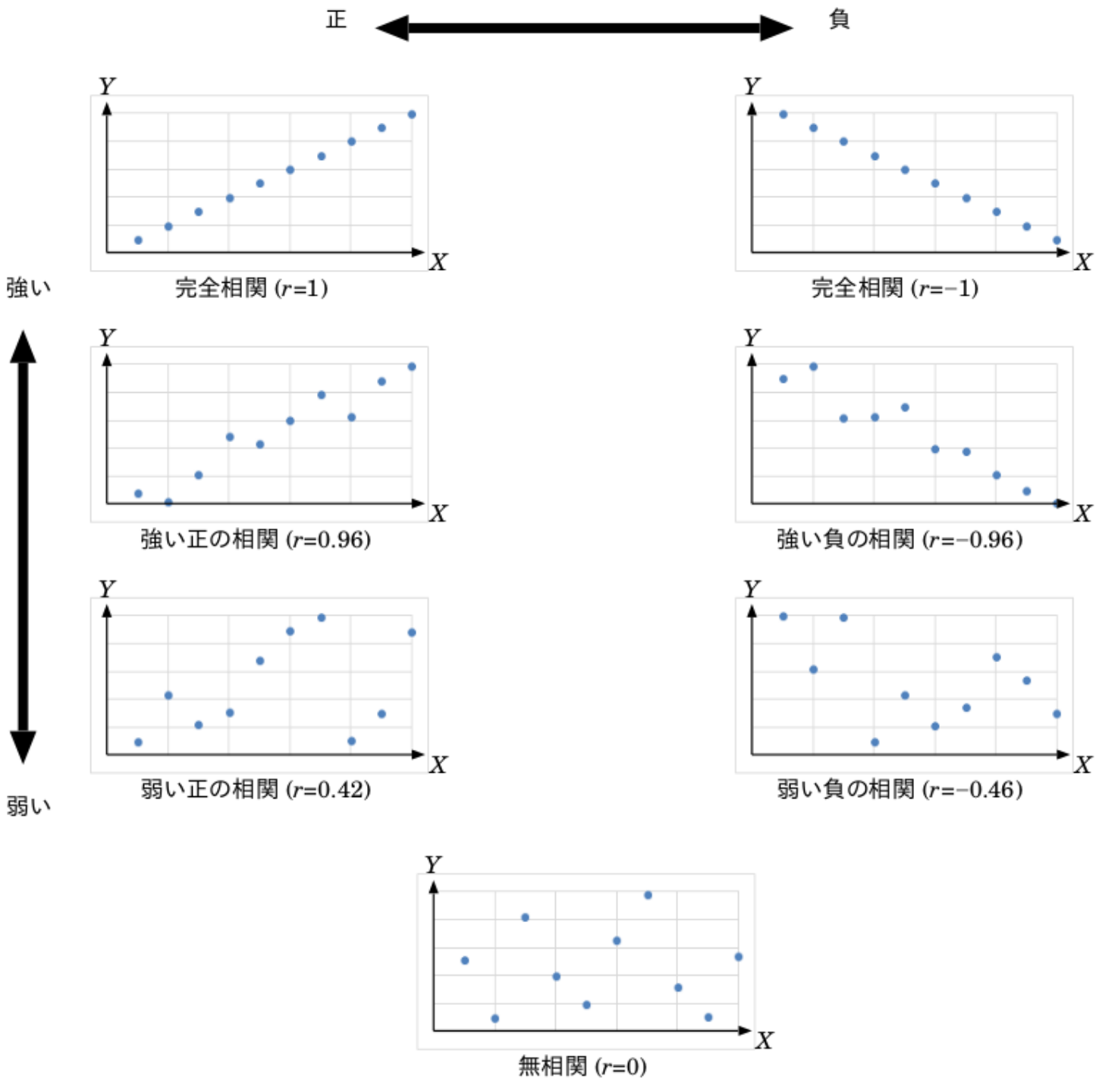
相関係数

相関関係を数量的に表す指標を**相関係数**という。

相関係数の性質

- −1から1までの値しかとらない。
- 値が正なら「正の相関」、負なら「負の相関」を表す。
- 値が±1に近いほど「強い相関」、0に近いほど「弱い相関」を表す。
- 値が±1なら「完全相関」、0なら「無相関」を表す。

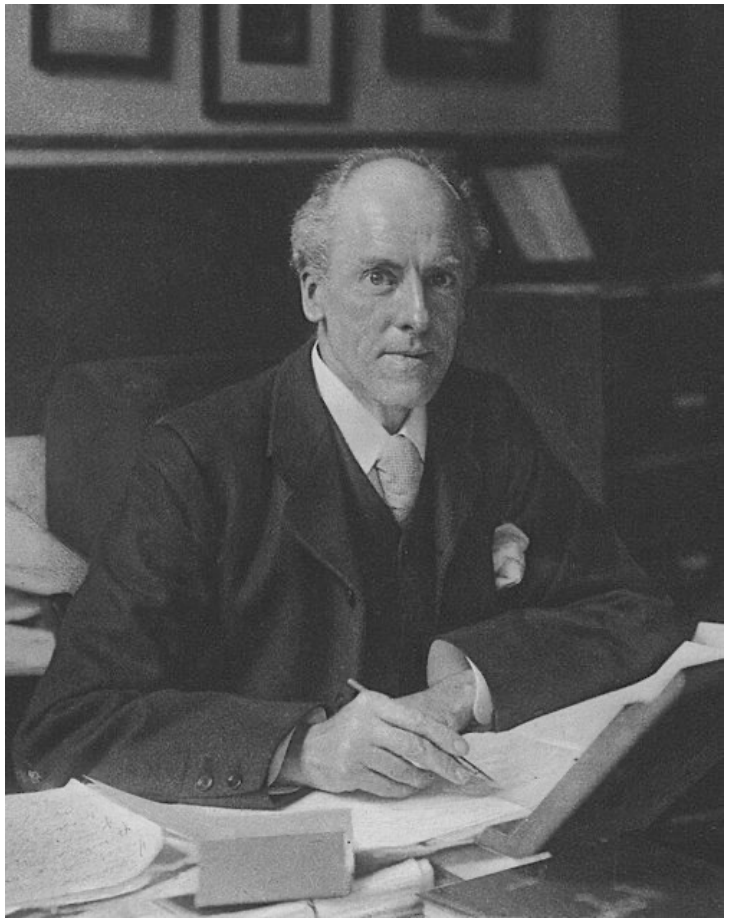
グラフの相関係数 *r* (ピアソンの積率相関係数)と相関関係の対応



相関係数

相関関係を数量的に表す指標を
相関係数という。

ピアソンの積率相関係数



カール・ピアソン

イギリスの統計学者カール・ピアソンが考案した相関係数で、単に「相関係数」といえばこれを指すことが多い。

ピアソンの積率相関係数

2変量のデータ $(X_i, Y_i)(i = 1,2, \ldots, n)$ に対するピアソンの積率相関係数rの定義は、

$$r = \frac{C_{XY}}{S_X S_Y}$$

ピアソンの積率相関係数

$$r = \frac{C_{XY}}{S_X S_Y}$$

ここで、C_{XY}は共分散という。

$$C_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

ピアソンの積率相関係数

$$r = \frac{C_{XY}}{S_X S_Y}$$

S_X, S_Y はそれぞれ X, Y の標準偏差である。

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad S_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

ピアソンの積率相関係数

$$r = \frac{C_{XY}}{S_X S_Y}$$

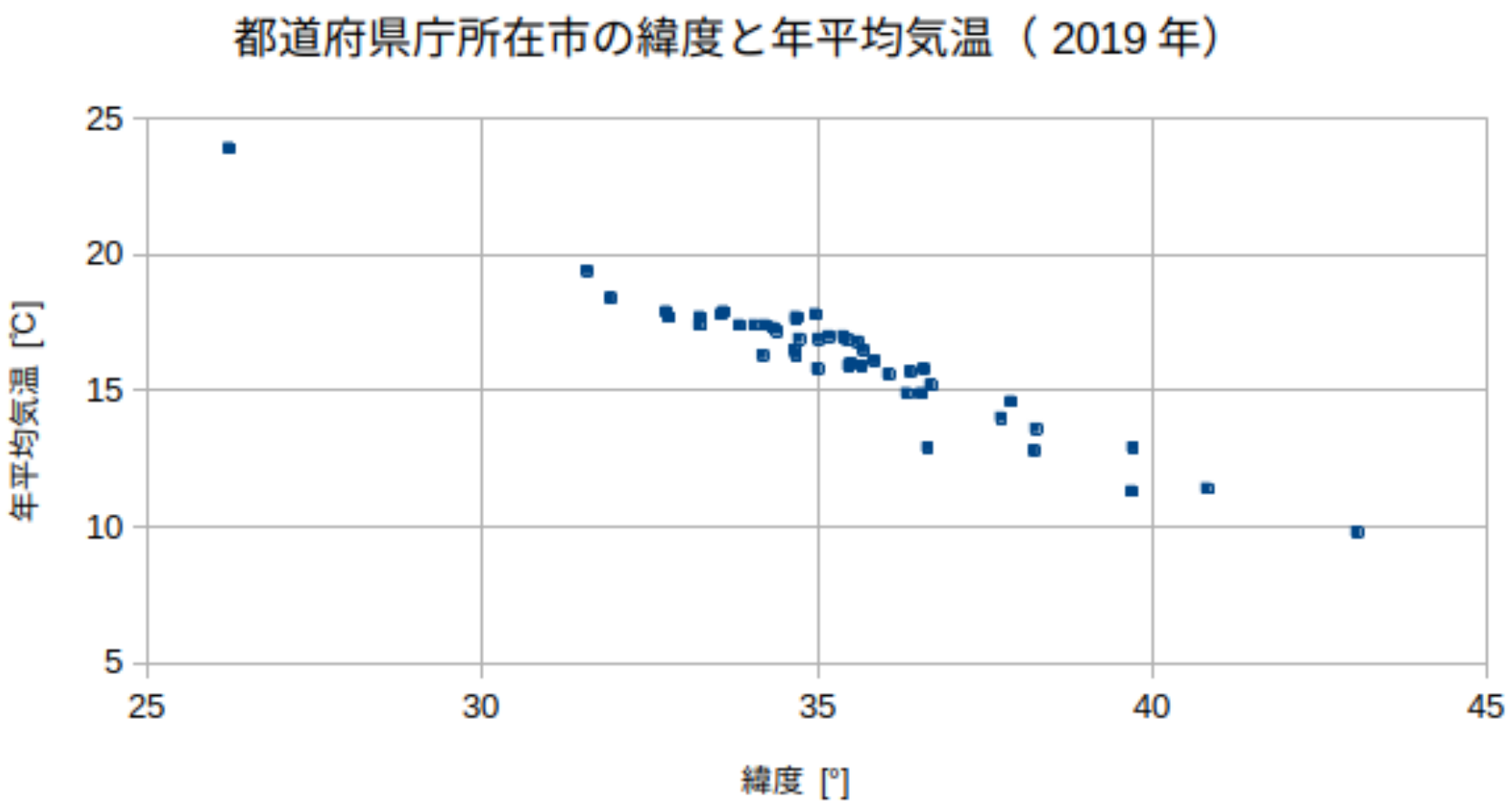
XとYは入れ替えても$C_{XY} = C_{YX}$なので、rは変わらない。

ピアソンの積率相関係数

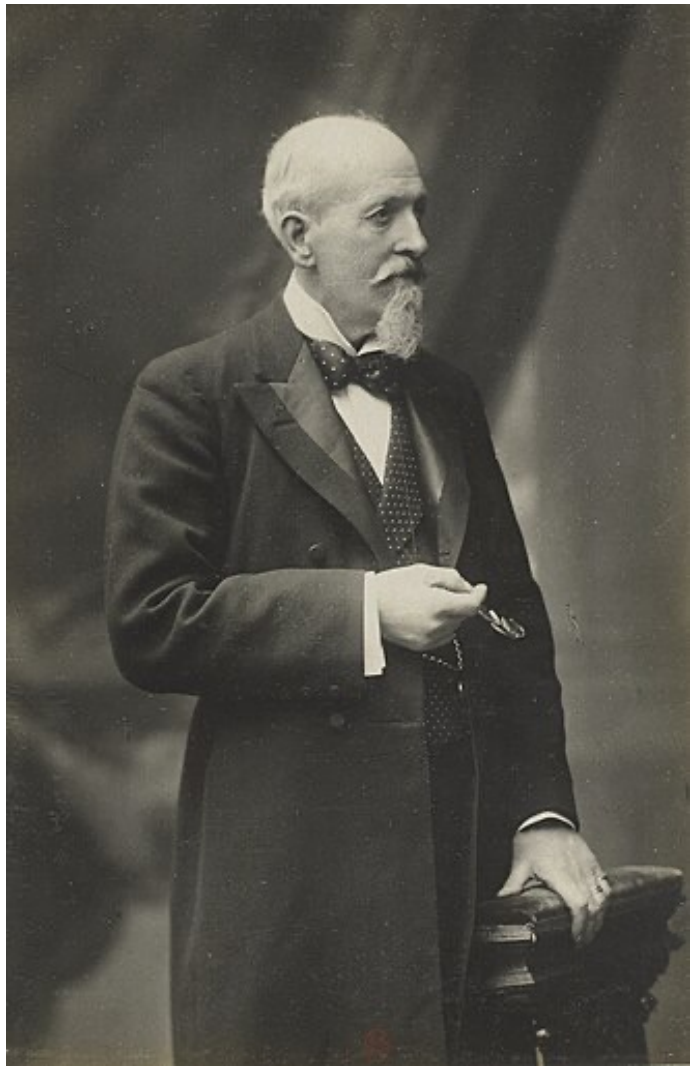
右図の場合の相関係数を求めると

$r = -0.962$

なので、「強い負の相関」を示している。



スピアマンの順位相関係数



チャールズ・スピアマン

イギリスの心理学者チャールズ・スピアマンが考案した相関係数。順位（順序尺度）データから求められる。順位をつければ順位のデータでなくても計算可能。

2変量の順位データ(R_i, R_i')

$(i = 1, 2, \cdots, n)$に対するスピアマンの順位相関係数r_sの定義は、

$$r_S = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (R_i - R_i')^2$$

同順位がある場合は修正（タイ修正）が必要。

相関行列

3変量以上の場合は、それぞれの変数から2変量の組み合わせを考え、それぞれの組み合わせで相関係数を求める。

この、それぞれの組み合わせで求めた相関係数を行列または表の形で表したものを**相関行列**という。

相関行列

例えば、*W, X, Y, Z*の4変量のデータの場合、2変量の組み合わせは

$${}_4C_2 = \frac{4!}{2!(4-2)!} = 6$$通り

あるため、相関係数は右の6つである。変数1と変数2は入れ替えても相関係数は変わらない。

変数1	変数2	相関係数
W	X	$r_{WX} = r_{XW}$
W	Y	$r_{WY} = r_{YW}$
W	Z	$r_{WZ} = r_{ZW}$
X	Y	$r_{XY} = r_{YX}$
X	Z	$r_{XZ} = r_{ZX}$
Y	Z	$r_{YZ} = r_{ZY}$

相関行列

- これを相関行列として表すと右のようになる。
- 対角線上にある相関係数は、同じ変量同士の相関係数なので、必ず1になる。
- 対角線の右上の相関係数は、対角線の左下の相関係数と同じなので省略されることが多い。

	W	X	Y	Z
W	1	r_{XW}	r_{YW}	r_{ZW}
X	r_{WX}	1	r_{YX}	r_{ZX}
Y	r_{WY}	r_{XY}	1	r_{ZY}
Z	r_{WZ}	r_{XZ}	r_{YZ}	1

相関係数の関数

関数	説明
CORREL(セル範囲1, セル範囲2)	セル範囲1 とセル範囲 2 のデータの相関係数（ピアソンの積率相関係数）を求める。

CORREL関数では1組の変数の相関係数（ピアソンの積率相関係数）を求めることができる。

今回の問い

- 「相関関係がある」とはどのような場合か？
  - 2変量が比例関係に近い傾向が見られる場合。
- **相関関係を数量的に表す指標とは何か？**
  - **相関係数**
- 相関関係のある2つの物事について解釈する際に気を付けるべきことは何か？

演習：相関行列

dataset1.odsから「年齢」「身長」「体重」の3変量の相関行列を求める。

演習：相関行列

1. 列Cから列Eまでを選択する（入力データ）。
2. メニューバーの[**データ**]→[**統計**]→[**相関**]を選択すると「相関」のダイアログが開く。
3. 「データ」の「結果貼り付け先」に結果を出力したい空のセルを指定する。
4. 「データ方向」はデータが列方向に並んでいるので[**列**]を選択する。
5. [**OK**]をクリックすると相関行列が出力される。
6. 出力された相関行列の項目は「列１」「列２」などになって分かりにくいので、入力データのラベルをコピー＆ペーストで貼り付けると良い。

この機能で用いる入力データは連続したセル範囲にならなければならない。

外れ値の影響

データの集団から大きく外れる値を**外れ値**という。

外れ値が生じる原因は、観測時の間違いやデータ入力の間違いなど様々である。何も間違いがないのに外れ値が生じることもある。

外れ値の影響

相関係数は外れ値によって大きく影響を受けることがあるので、散布図を描いて確認してみることが必要である。

外れ値の影響

外れ値が生じた場合は、その原因を調査して間違いであれば修正したり取り除いたりする必要がある。

外れ値の影響

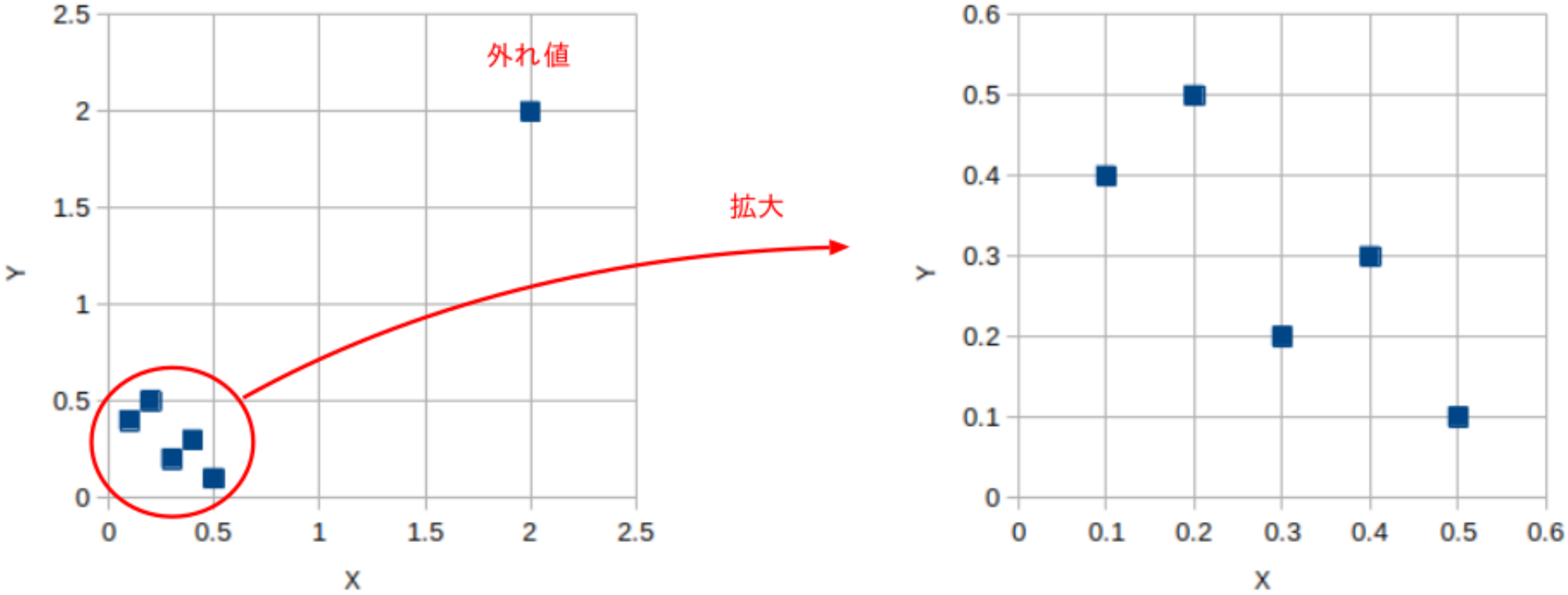
例として右のデータを考える。

このデータのピアソンの積率相関係数はr=0.928であり、これだけを見ると強い正の相関がありそうである。

番号	X	Y
1	0.2	0.5
2	0.1	0.4
3	0.4	0.3
4	0.5	0.1
5	0.3	0.2
6	2	2

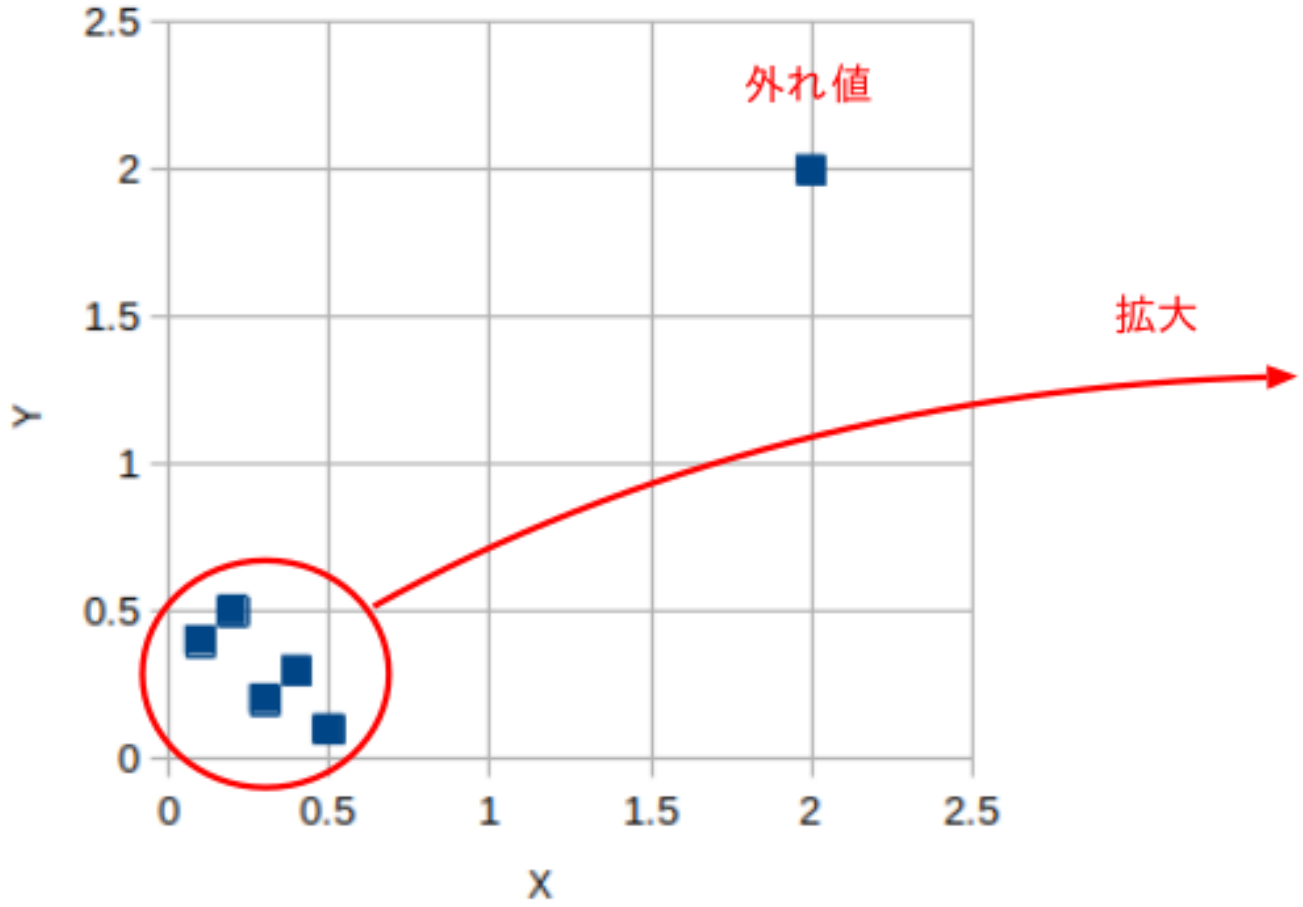
外れ値の影響

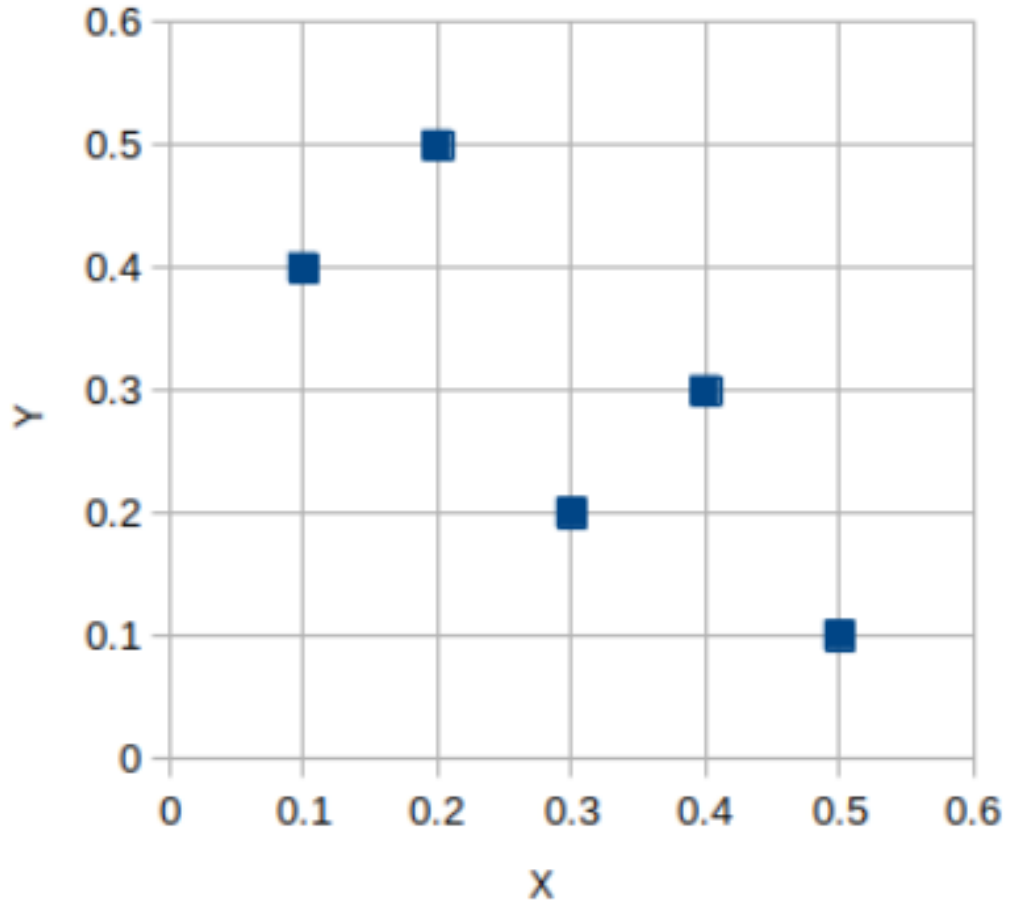
しかし、散布図を描くと６番目のデータはその他のデータと比べて大きく異なり、外れ値であることが分かる。



外れ値の影響

この外れ値を除いて1から5番目までのデータでピアソンの相関係数を計算すると $r = -0.8$ で、先ほどと真逆の相関関係になる。





今回学ぶこと

1. 相関関係
2. 相関係数
3. **因果関係と相関関係**

因果関係

2つの物事が原因と結果の関係にあること。原因の方を変化させるとそれに応じて結果も変化する。

相関関係

2つの物事の間に（直線的な）関係性があること。どちらが原因でどちらが結果かは分からない。

因果関係と相関関係

相関関係があるからといって無条件に因果関係があるとはいえない。

因果関係と相関関係

2つの物事の間に相関関係が見られた場合に、その2つがどういう関係で増えたり減ったりするのかをよく考えて解釈しなければならない。

因果関係の逆転

気温とTシャツの売り上げの間には正の相関関係がある。

因果関係の逆転

気温の上昇とともにTシャツの売り上げが増えるのは、暑くなるとみんな薄着になるから。

因果関係の逆転

因果関係：

原因：気温

結果：Tシャツの売り上げ

因果関係の逆転

この因果関係を逆に考えて、Tシャツを売れなくすれば気温が下がるかというと、それはありえない。

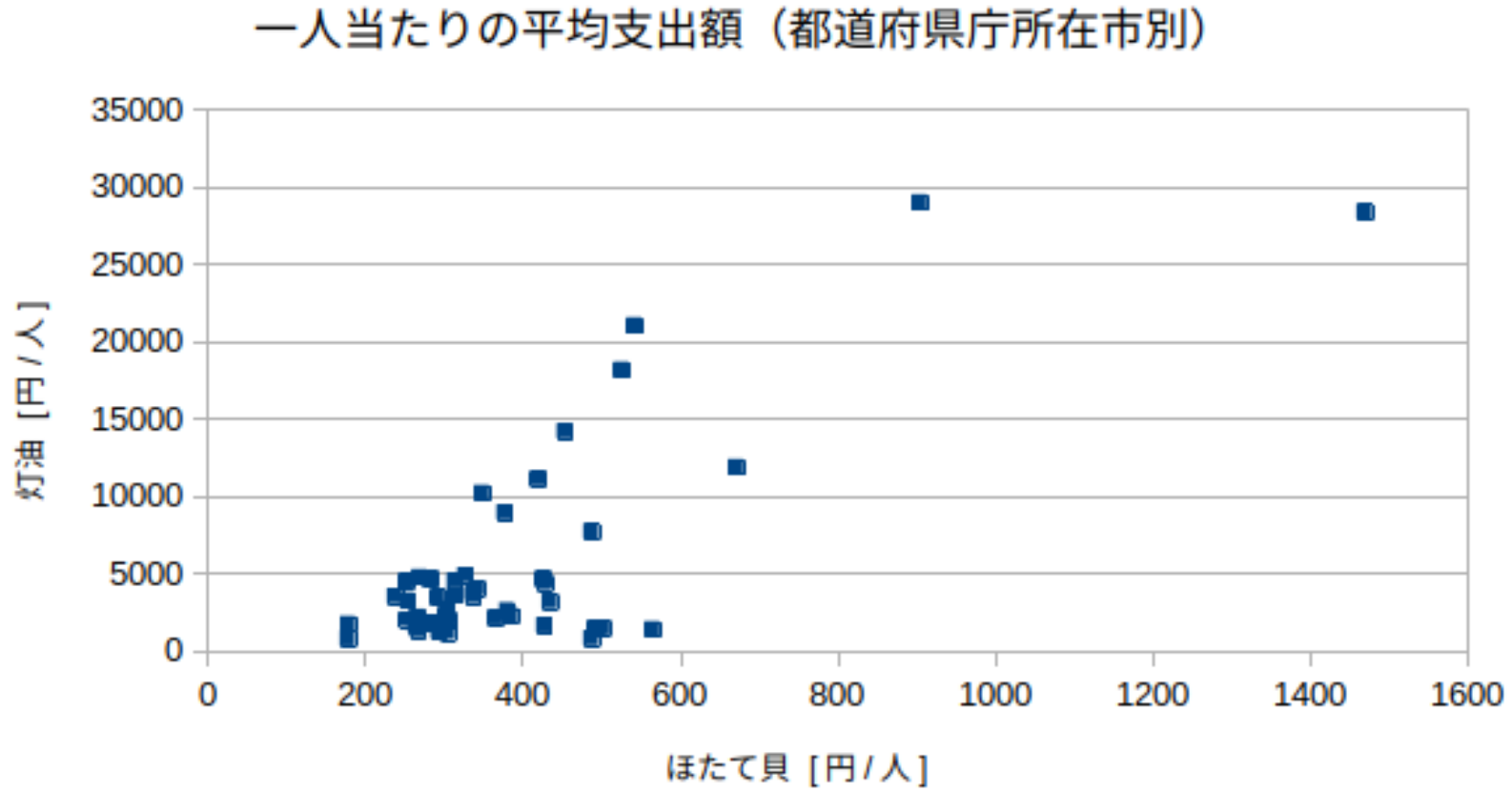
擬似相関

家計調査：

総務省統計局が全国の世帯を対象に家計収支の実態を把握するために行っている標本調査。

擬似相関

以下の図は、2020年家計調査における全国の都道府県庁所在市の品目別支出額の総世帯平均より、ほたて貝と灯油の一人当たりの平均支出額を算出し、散布図にしたものである。



出典：総務省統計局「家計調査年報（家計収支編）2020年（令和2年）」（参照2021-09-09）

擬似相関

ほたて貝への支出額が多いと灯油への支出額も多いという正の相関が見られる ($r = 0.778$)。しかし、ほたて貝と灯油が直接何かに関係しているようには思えない。

擬似相関とは

一見関係なさそうなことの間に相関関係が見られること。

擬似相関

ほたて貝と灯油の支出で共通しているのは、緯度の高い市ほど支出額が多く、緯度の低い市ほど支出額が少ないということ。

都道府県庁所在市	都道府県庁の緯度 [°]	ほたて貝の支出額 [円/人]	灯油の支出額 [円/人]
札幌市	43.064	904.4	28974.6
青森市	40.084	1469.6	28349.4
盛岡市	39.704	526.0	18159.3
仙台市	38.269	671.2	11867.7
⋮	⋮	⋮	⋮
鹿児島市	31.560	263.9	1693.1
沖縄市	26.213	177.9	787.7

擬似相関

ほたて貝は水温が低いところを好むため、北の方ほど養殖が盛んで消費も多い。

また、北の方ほど冬は寒いので暖房に灯油を使う量が多くなる。

擬似相関

緯度がほたて貝と灯油の支出額のどちらにも影響しているように思われる。

この例の緯度のような擬似相関における隠れた要因を**交絡因子**という。

今回の問い

- 「相関関係がある」とはどのような場合か？
  - 2変量が比例関係に近い傾向が見られる場合。
- 相関関係を数量的に表す指標とは何か？
  - 相関係数
- **相関関係のある2つの物事について解釈する際に気を付けるべきことは何か？**
  - **その２つがどういう関係で増えたり減ったりするのかをよく考える必要がある。**

今回学んだこと

1. 相関関係
2. 相関係数
3. 因果関係と相関関係

今回の目標

2変量データの相関について定量的に分析できるようになる。

課題

dataset2.odsの各科目間の相関行列を求めなさい。全部で6科目あるので6変量のデータである。

作成したファイルはODF表計算ドキュメント(拡張子.ods)で提出すること。

提出先：yutaka.hirai@koeki-u.ac.jp

件名：データリテラシー第10回課題 [学籍番号]

締め切り：12月18日 (金)

次回

第１回　社会におけるデータ・AI利活用
第２回　表計算ソフトの基本操作
第３回　セルの参照と集計
第４回　度数分布表とヒストグラム
第５回　代表値と散らばりの尺度
第６回　データのグラフによる表現とデータ比較
第７回　データの分布と標本抽出
第８回　推定
第９回　分割表とクロス集計
第１０回　２変量データと相関
第１１回　時系列データ
第１２回　仮説検定
第１３回　回帰分析