

# 社会人のためのデータサイエンス演習

第3週：分析の具体的手法

第1回：クロス集計の軸設定と見方

講師名：大黒 健一

# 講義内容

## 第1週

- データサイエンスとは

## 第2週

- 分析の概念と事例  
ビジネス課題解決のためのデータ分析基礎(事例と手法)①

## 第3週

- 分析の具体的手法  
ビジネス課題解決のためのデータ分析基礎(事例と手法)②

## 第4週

- ビジネスにおける予測と分析結果の報告  
ビジネス課題解決のためのデータ分析基礎(事例と手法)③

## 第5週

- ビジネスでデータサイエンスを実現するために

# 第3週の内容紹介

---

## 第1回

- クロス集計の軸設定と見方

## 第2回

- 散布図と相関の調べ方

## 第3回

- 相関関係と因果関係の違い

## 第4回

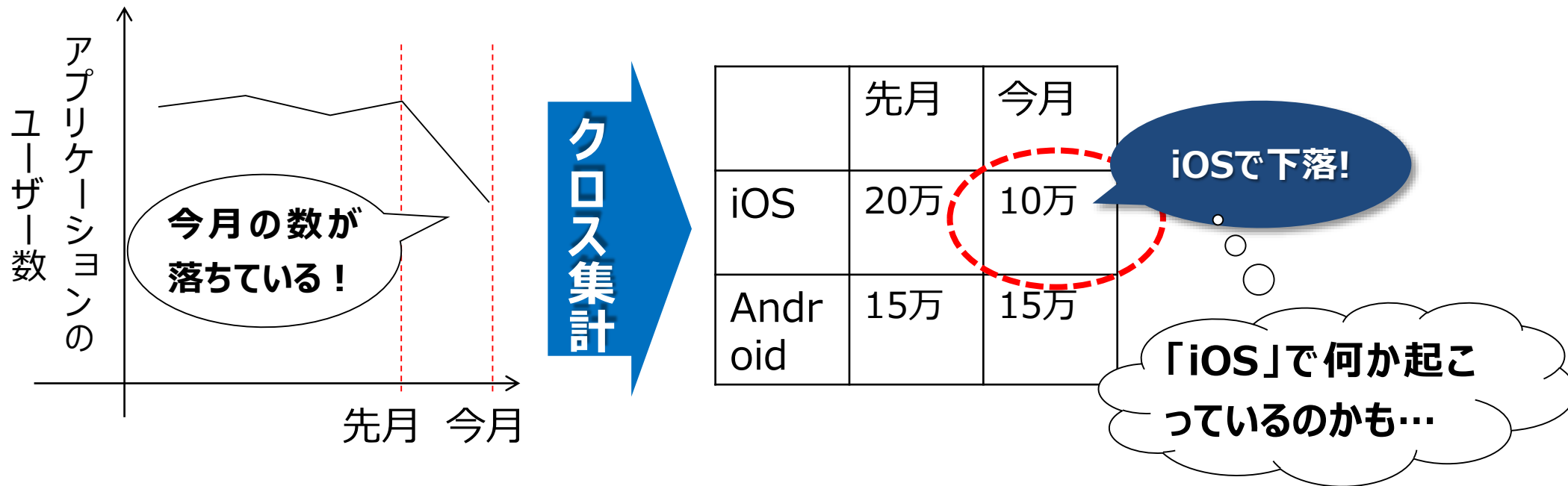
- 時系列データの見方

## 第5回

- 時系列データの分解の方法

# クロス集計とは

- 2変数のカテゴリの組み合わせについてデータ個数や比率を集計
- 横カテゴリと縦カテゴリの関連を調査可能
  - 変数間の関連性が明らかになると、課題や注力ポイントが発見しやすくなる。



**目的に応じた軸(切り口)を設定することが重要**

# クロス集計の軸設定の考え方

## ● 軸となる変数は、カテゴリーデータ

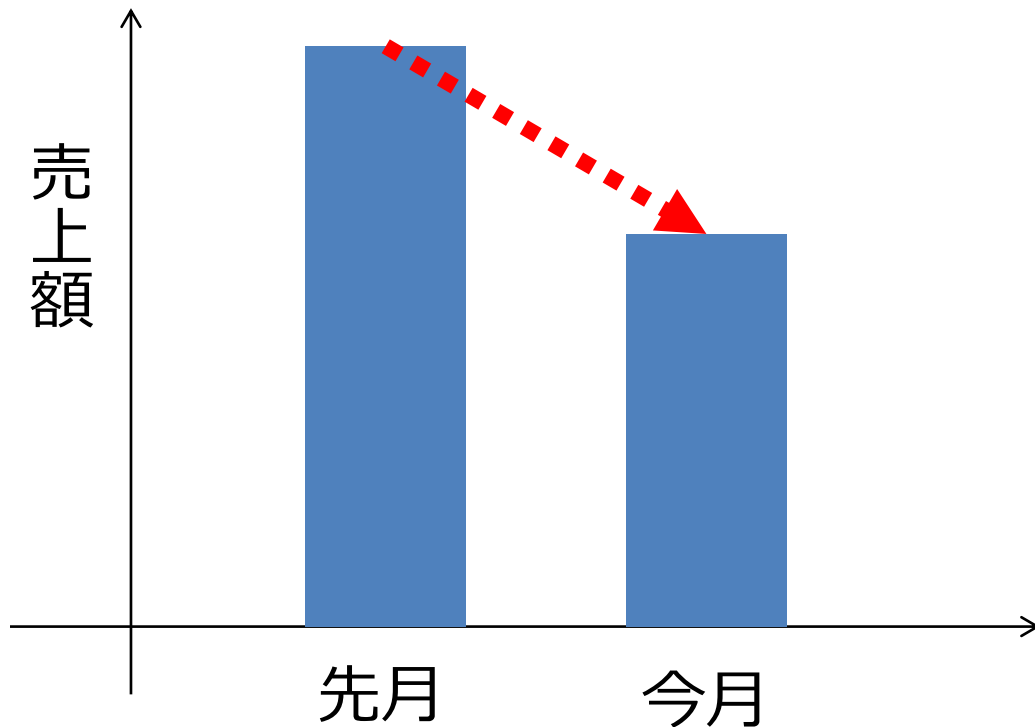
- 関係性を把握したい変数を軸に取る

### ■ 変数の例

| 分類         | 特徴                                         | 変数の例                  |
|------------|--------------------------------------------|-----------------------|
| デモグラフィック変数 | 人口統計分布に基づく変数。基礎的情報として見ることが多い。              | 性別、年代、家族構成、職業、年収、学歴など |
| 地理的変数      | 地理的に分割される変数。消費者向けのマーケティングで利用されることが多い。      | 国、行政区、気候地域、都市と農村など    |
| 心理的変数      | 価値観やライフスタイル、好み等を表す変数。意識調査の結果から変数をつくることが多い。 | * * が好き、* * を常用しているなど |
| 行動変数       | 行動を表す変数。IT化の進展によって集計が容易になった。               | 購買履歴、使用頻度、アクセスログなど    |

# 問題

- グラフは、ある小売り店舗の売上額を先月と今月で比較したもの。
- 今月の売上が落ち込んでいます。
- 何に問題があるのでしょうか。



# クロス軸を検討する

---

## ● 売上を低下させる要素の仮説を出す

- 夏休みが終わったので、ファミリー層の売上が減ったのではないかな？  
⇒ ファミリー層の売上が落ちている
- 雨の日が多かったなので、天候のせいではないかな？  
⇒ 雨の日は、晴れや曇りの日に比べて売上が落ちる

## ● 仮説にもとづいて軸を設定する

- ファミリー層に着目 ⇒ 「家族構成」という変数を軸に設定してみる
- 天候に着目 ⇒ 「天気」という変数を軸に設定してみる

# クロス集計表から読み取る①

## ● クロス集計表を作成し、解釈する

➤ 家族構成(ファミリー層、単身層)を軸に設定

| 上段：n<br>下段：% | 先月              | 今月              |                  |
|--------------|-----------------|-----------------|------------------|
| ファミリー層       | 200,000円<br>67% | 100,000円<br>33% | 300,000円<br>100% |
| 単身層          | 180,000円<br>50% | 180,000円<br>50% | 360,000円<br>100% |
| 計            | 380,000円<br>58% | 280,000円<br>42% | 660,000円<br>100% |

ファミリー層で  
売上が落ちている  
ことが判明

表側

行パーセント

※列方向は「列パーセント」

# クロス集計表から読み取る②

## ● 異なる軸を設定すると・・・

■「家族構成」を軸に設定

|        | 先月              | 今月              | 計                |
|--------|-----------------|-----------------|------------------|
| ファミリー層 | 200,000円<br>67% | 100,000円<br>33% | 300,000円<br>100% |
| 単身層    | 180,000円<br>50% | 180,000円<br>50% | 360,000円<br>100% |
| 計      | 380,000円<br>58% | 280,000円<br>42% | 660,000円<br>100% |

差異が明確

■「天候」を軸に設定

|       | 先月              | 今月              | 計                |
|-------|-----------------|-----------------|------------------|
| 雨の日   | 110,000円<br>58% | 80,000円<br>42%  | 190,000円<br>100% |
| 雨以外の日 | 270,000円<br>57% | 200,000円<br>43% | 470,000円<br>100% |
| 計     | 380,000円<br>58% | 280,000円<br>42% | 660,000円<br>100% |

売上にはあまり関係がない

様々な軸でクロス分析を行うことで状況の詳細な把握が可能

# 次回のテーマ

---

次回は

**「散布図と相関の調べ方」**

お疲れ様でした！

# 社会人のためのデータサイエンス演習

第3週：分析の具体的手法

第2回：散布図と相関の調べ方

講師名：大黒 健一

# 第3週の内容紹介

---

## 第1回

- クロス集計の軸設定と見方

## 第2回

- **散布図と相関の調べ方**

## 第3回

- 相関関係と因果関係の違い

## 第4回

- 時系列データの見方

## 第5回

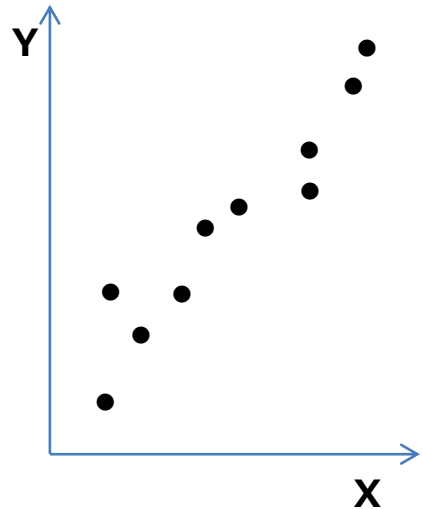
- 時系列データの分解の方法

# 散布図と相関

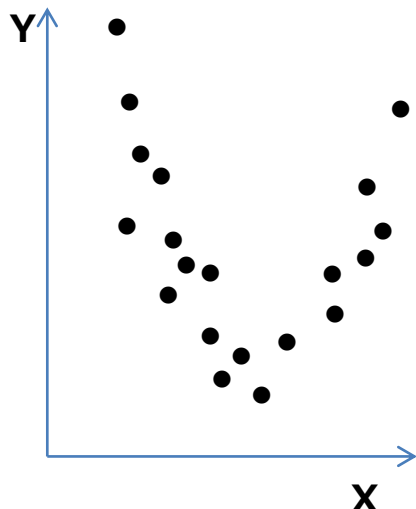
## ● 2変数の関係性を見るには、散布図が有効

- 散布図は、2つの変数をそれぞれX軸、Y軸として、1レコードごとの座標を点で示す。
- 下図Aのような線形の関係だけでなく、B非線形の関係、C離散的な関係、D外れ値を含む関係のように、様々な関係性を観察することができる。

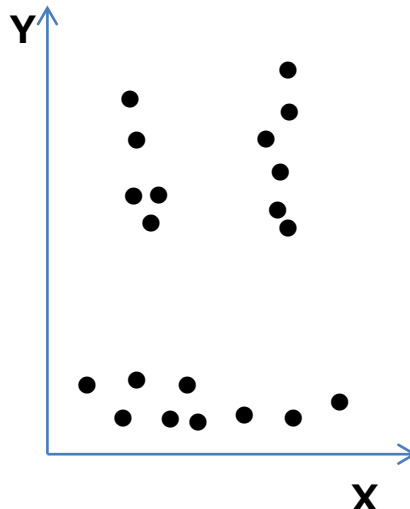
A 線形の関係



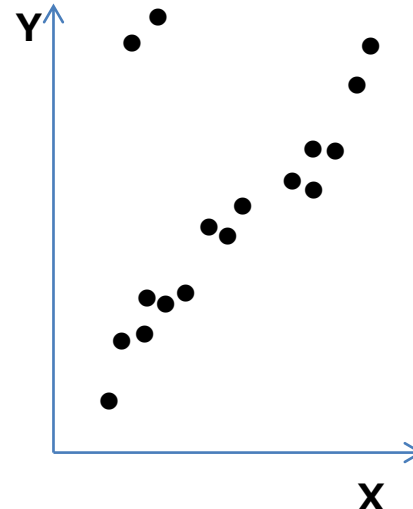
B 非線形の関係



C 離散的な関係



D 外れ値を含む



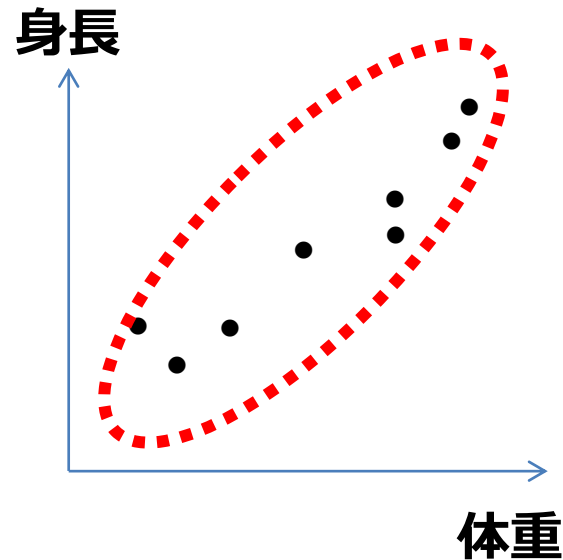
プロットすると、関係が見えてくる

# 相関とは

- 線形の関係において、2つの変数に関係があることを「相関」という

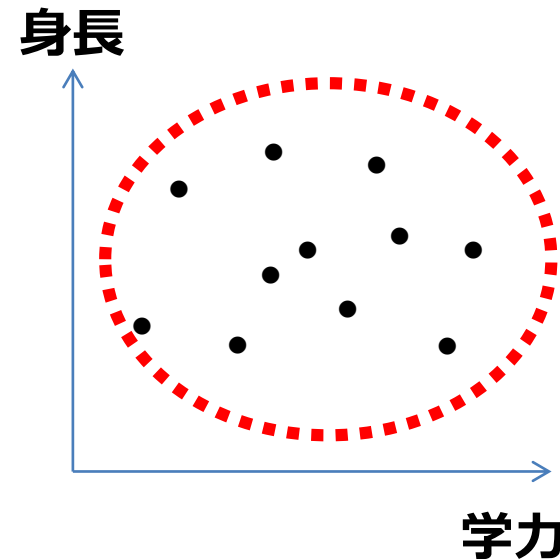
ある変数が増大すればするほど、もう一方の変数が増大

正の相関



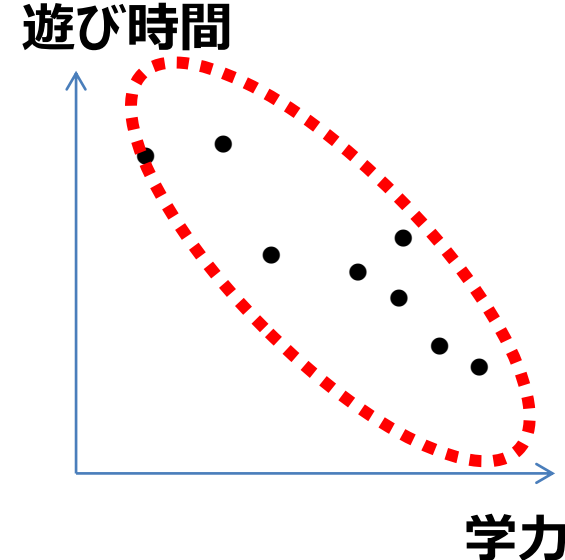
ある変数が増大しても、もう一方の変数は無関係な値

相関なし



ある変数が増大すればするほど、もう一方の変数が減少

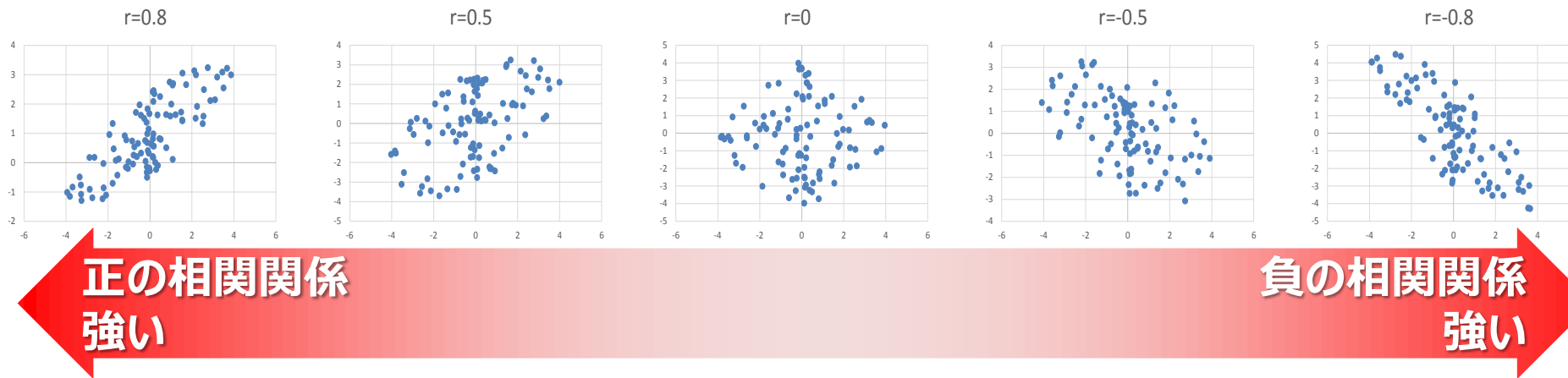
負の相関



# 相関係数と散布図の関係

- 2つの変数の関係に関する基本統計量のことを、**相関係数( $r$ )**という

➤ 相関の強さの絶対的な基準はないが、一般的には以下のように捉えられる。



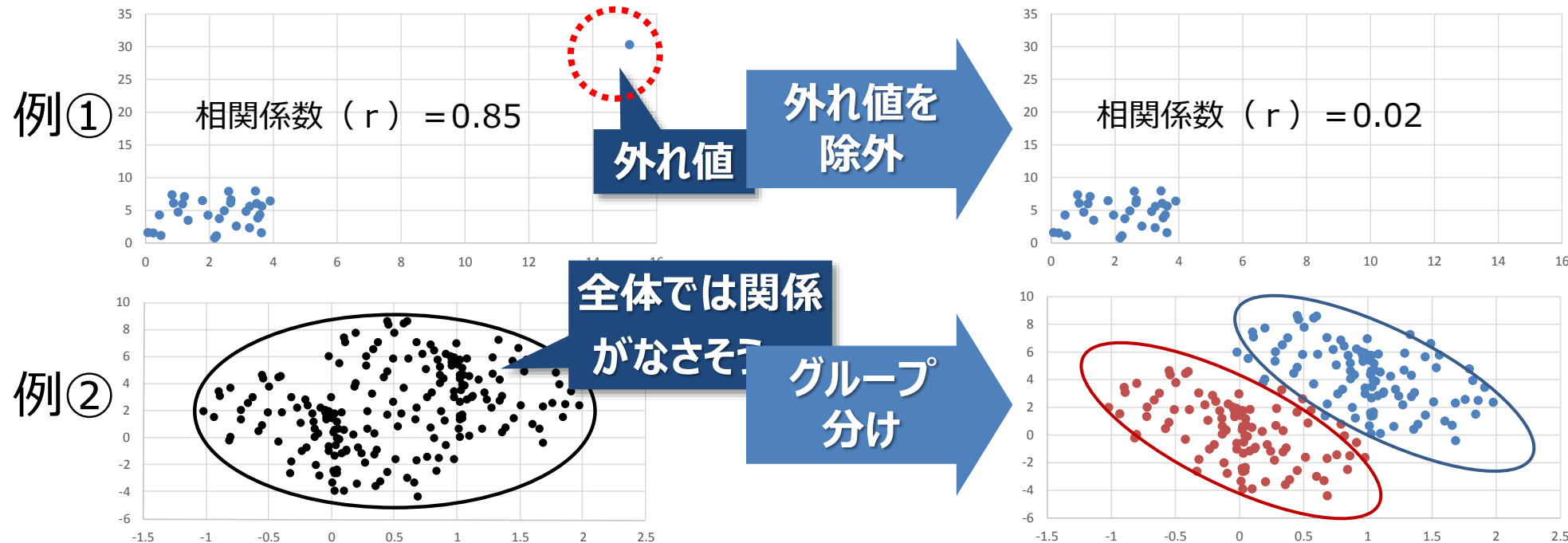
| 正の相関    | 解釈       |
|---------|----------|
| 0～0.1   | 無相関      |
| 0.1～0.3 | 弱い正の相関   |
| 0.3～0.7 | 中程度の正の相関 |
| 0.7～1   | 強い正の相関   |

| 負の相関      | 解釈       |
|-----------|----------|
| -0.1～0    | 無相関      |
| -0.3～-0.1 | 弱い負の相関   |
| -0.7～-0.3 | 中程度の負の相関 |
| -1～-0.7   | 強い負の相関   |

# 相関係数を算出する前に

## ● 散布図を描く

- 散布図を描かずに、いきなり相関係数を出して解釈すると、
- 外れ値に気づかず、相関が高いと判断…
- 無相関と思っていたのに、グループ分けによって相関が発現…



**相関係数の値を過信せず、  
散布図を注意深く観察することが必要**

# 問題

- 下記は水稲の作付面積と収穫量データです。
- 作付面積と収穫量に係り性はあるでしょうか。

| 年次    | 作付面積(千ha) | 収穫量(千 t ) |
|-------|-----------|-----------|
| 1990年 | 2,055     | 10,463    |
| 1991年 | 2,033     | 9,565     |
| 1992年 | 2,092     | 10,546    |
| 1993年 | 2,127     | 7,811     |
| 1994年 | 2,200     | 11,961    |
| 1995年 | 2,106     | 10,724    |
| 1996年 | 1,967     | 10,328    |
| 1997年 | 1,944     | 10,004    |
| 1998年 | 1,793     | 8,939     |
| 1999年 | 1,780     | 9,159     |
| 2000年 | 1,763     | 9,472     |
| 2001年 | 1,700     | 9,048     |
| 2002年 | 1,683     | 8,876     |
| 2003年 | 1,660     | 7,779     |
| 2004年 | 1,697     | 8,721     |

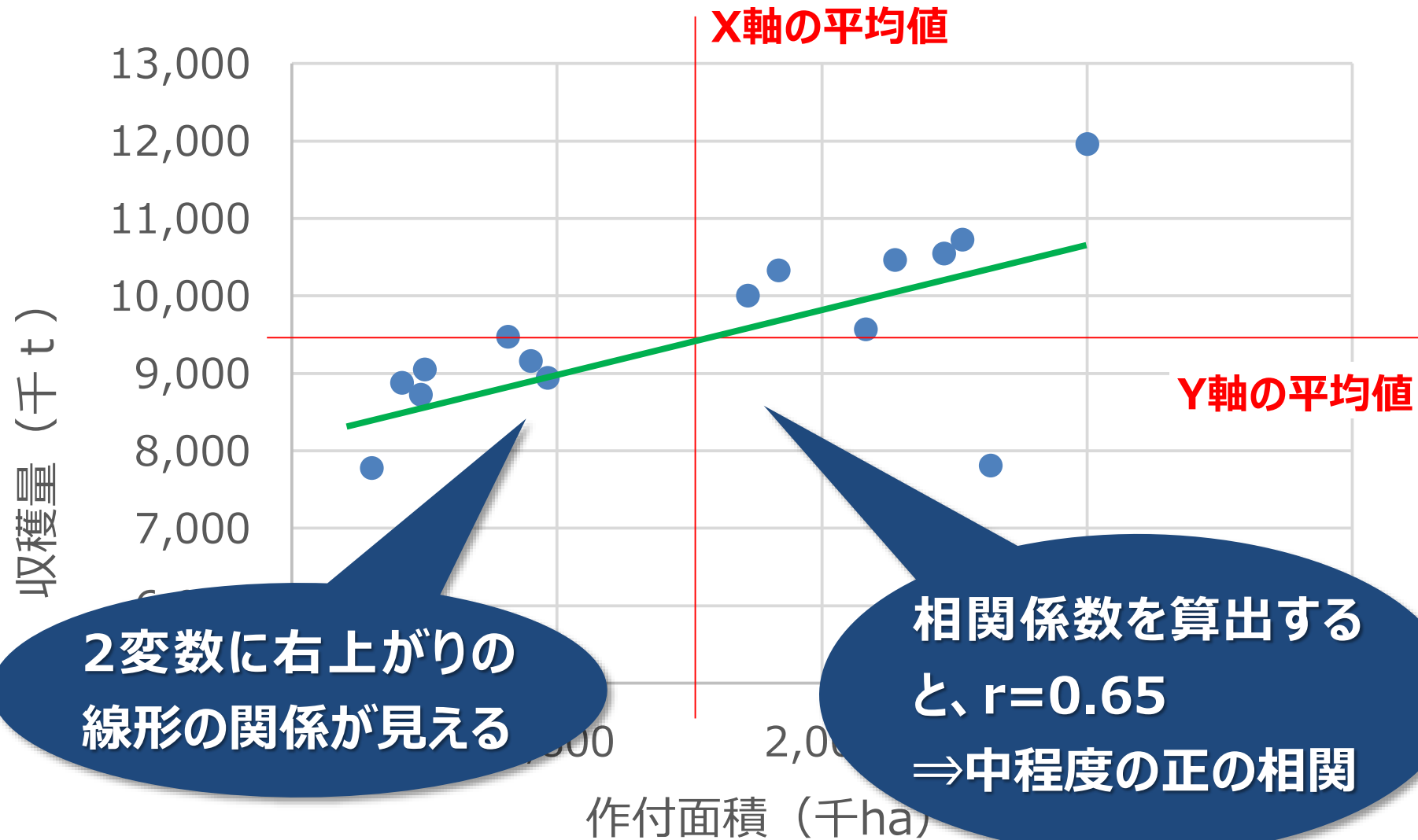
|      | 作付面積   | 収穫量       |
|------|--------|-----------|
| 平均値  | 1,907  | 9,560     |
| 分散   | 35,630 | 1,279,525 |
| 標準偏差 | 188.8  | 1131.2    |

出典：農林水産省 作物統計より作成

[http://www.e-stat.go.jp/SG1/estat/GL08020103.do?\\_toGL08020103\\_&tcID=000001024932&cycleCode=0&requestSender=dsearch](http://www.e-stat.go.jp/SG1/estat/GL08020103.do?_toGL08020103_&tcID=000001024932&cycleCode=0&requestSender=dsearch)

# 相関関係を調べる①

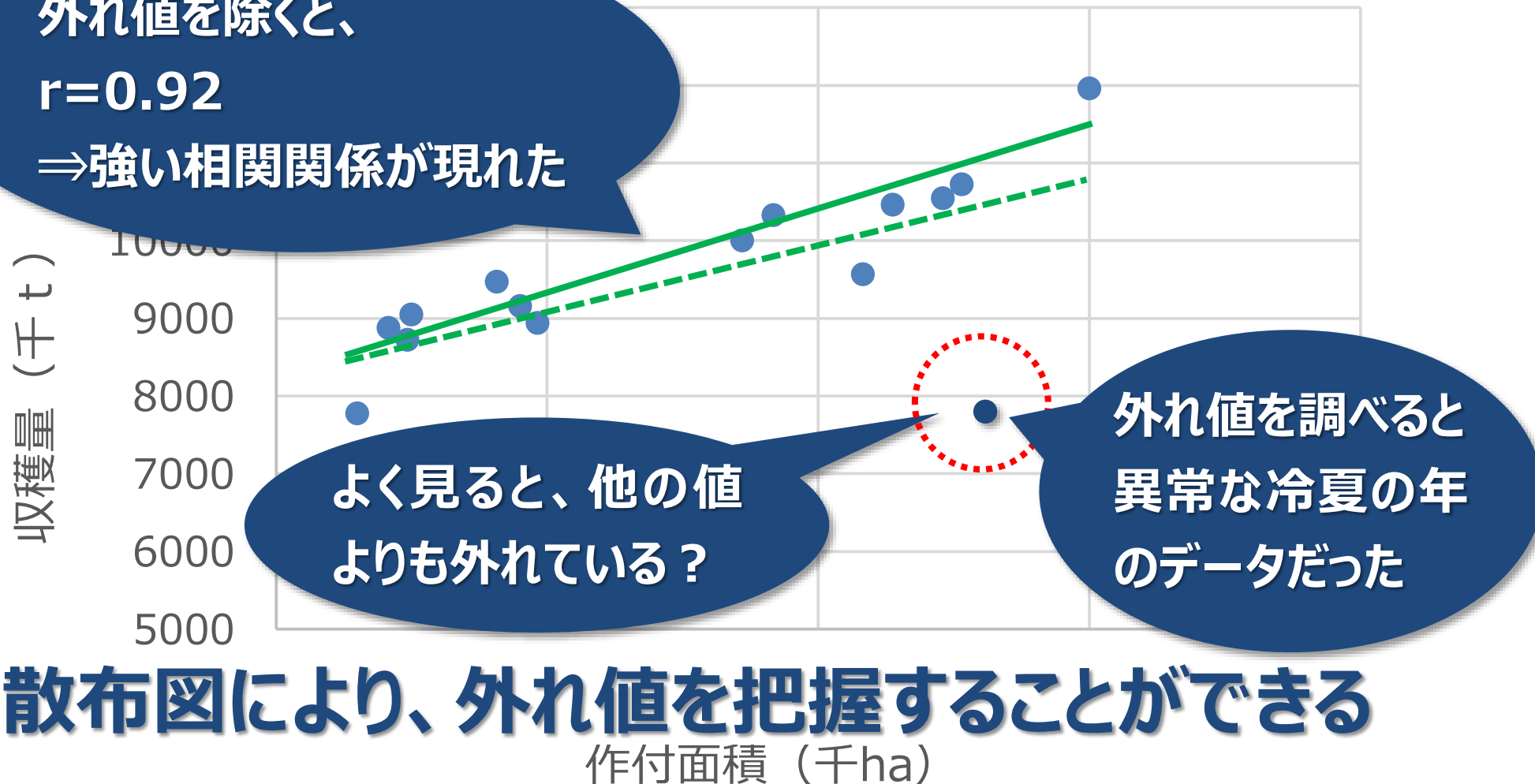
- 散布図を描いてみる。



# 相関関係を調べる②

## ● これで分析は終わり？

外れ値を除くと、  
 $r=0.92$   
⇒強い相関関係が現れた



散布図により、外れ値を把握することができる

# 相関係数を調べる③

## ● Excelで相関係数を算出するには

- 左図のような2変数の相関係数を算出する場合、エクセル関数の「CORREL」で算出できる。
- 値を出したいセルに、
- **=CORREL(変量1、変量2)**
- と記入

|    | A     | B                      | C           | D |
|----|-------|------------------------|-------------|---|
|    | 年次    | 作付面積<br>(千ha)          | 収穫量<br>(千t) |   |
| 1  |       |                        |             |   |
| 2  | 1990年 | 2,055                  | 10,463      |   |
| 3  | 1991年 | 2,033                  | 9,565       |   |
| 4  | 1992年 | 2,092                  | 10,546      |   |
| 5  | 1993年 | 2,127                  | 7,811       |   |
| 6  | 1994年 | 2,200                  | 11,961      |   |
| 7  | 1995年 | 2,106                  | 10,724      |   |
| 8  | 1996年 | 1,967                  | 10,328      |   |
| 9  | 1997年 | 1,944                  | 10,004      |   |
| 10 | 1998年 | 1,793                  | 8,939       |   |
| 11 | 1999年 | 1,780                  | 9,159       |   |
| 12 | 2000年 | 1,763                  | 9,472       |   |
| 13 | 2001年 | 1,700                  | 9,048       |   |
| 14 | 2002年 | 1,683                  | 8,876       |   |
| 15 | 2003年 | 1,660                  | 7,779       |   |
| 16 | 2004年 | 1,697                  | 8,721       |   |
| 17 |       |                        |             |   |
| 18 | 相関係数  | =CORREL(B2:B16,C2:C16) |             |   |

# 次回のテーマ

---

次回は

**「相関関係と因果関係の違い」**

お疲れ様でした！

# 社会人のためのデータサイエンス演習

第3週：分析の具体的手法

第3回：相関関係と因果関係の違い

講師名：大黒 健一

# 第3週の内容紹介

---

## 第1回

- クロス集計の軸設定と見方

## 第2回

- 散布図と相関の調べ方

## 第3回

- **相関関係と因果関係の違い**

## 第4回

- 時系列データの見方

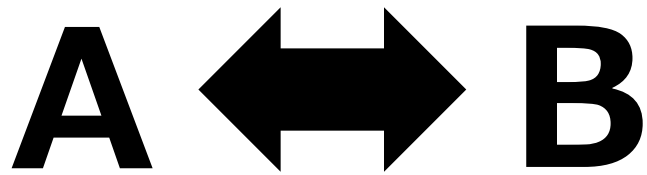
## 第5回

- 時系列データの分解の方法

# 相関と因果の違い

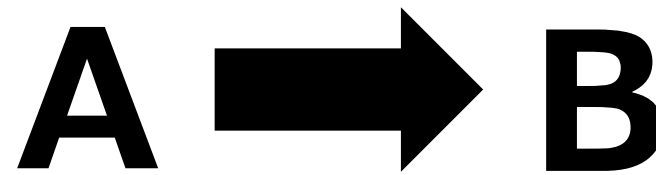
- **相関**は、ある変数が変化すると、他方の変数も同時に変化する関係
- **因果**は、ある変数が、他方の変化を引き起こす関係(一方通行の関係、原因と結果)

**相関関係**



背が高い人ほど体重も重い

**因果関係**



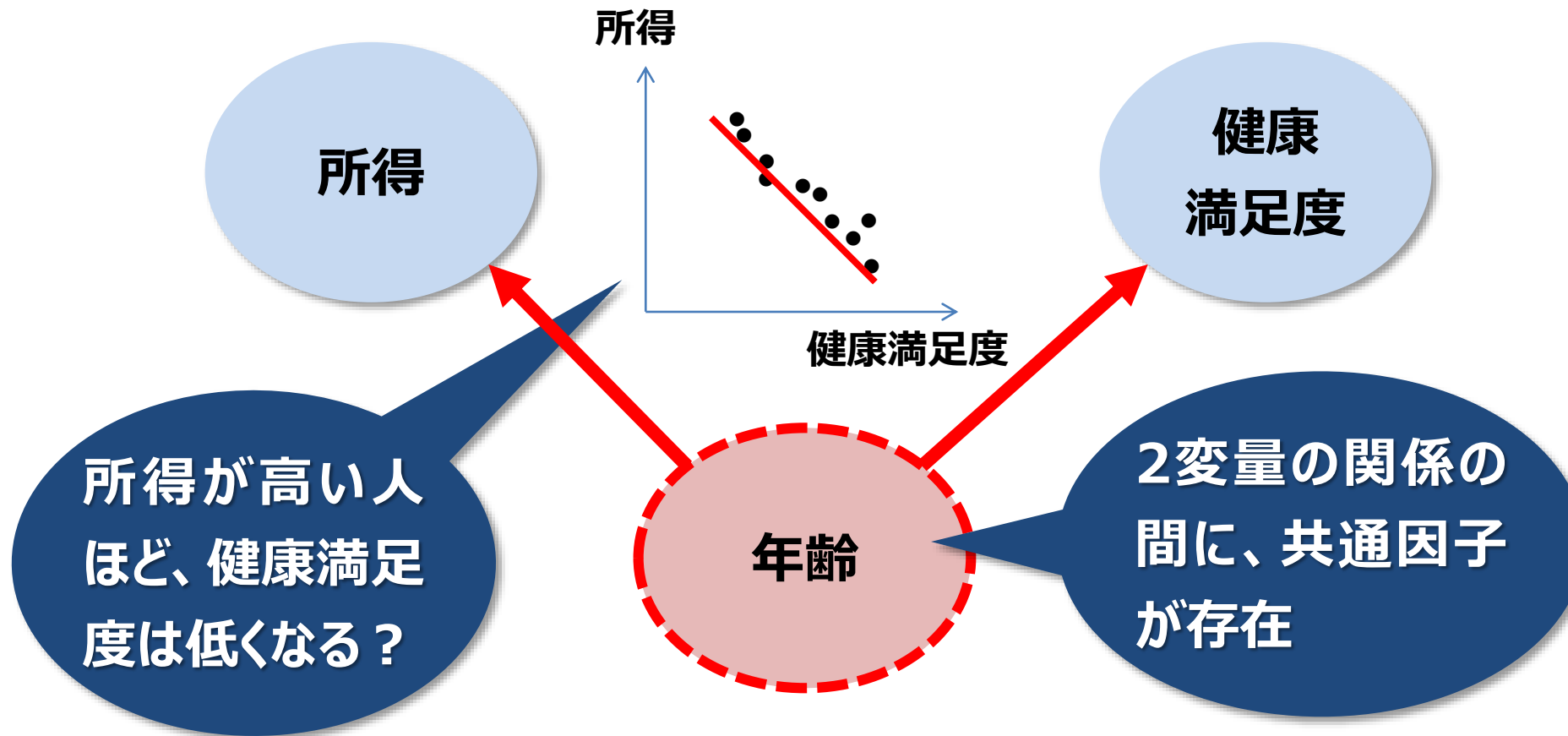
台風で航空便が欠航

**相関と因果は異なることに注意**

# 見せかけの相関関係が現れるケース①

## ● 因果の間に、共通の要因がある

- 例えば、所得と健康満足度の間には、負の相関関係がみられる。

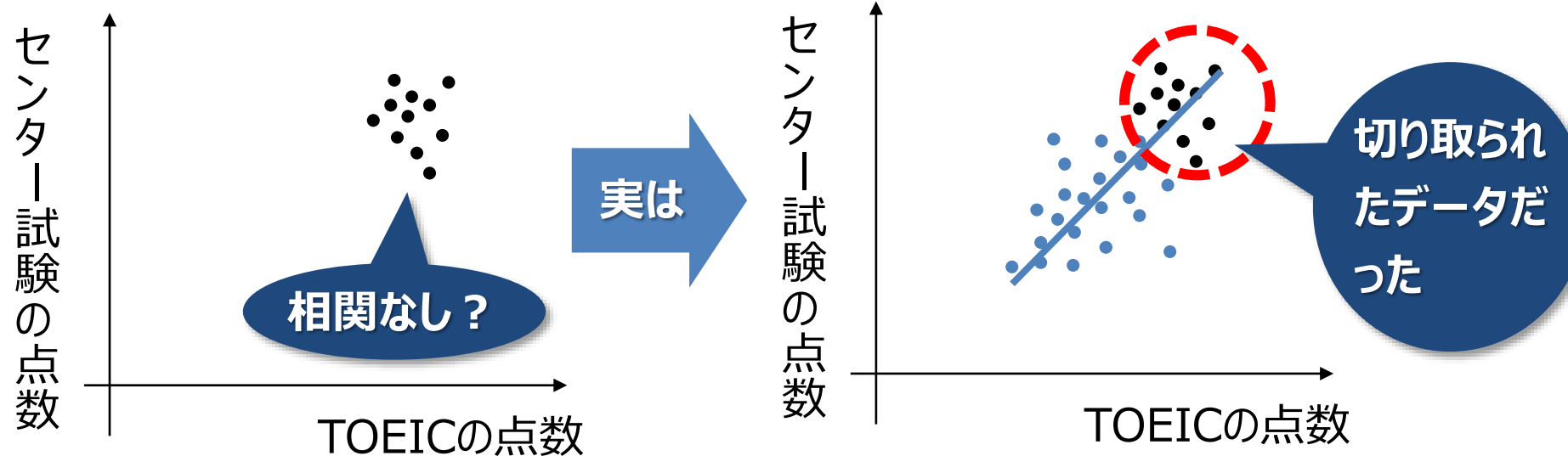


背景にある「共通因子」に注意

# 見せかけの相関関係が現れるケース②

## ● データのバイアス

- 例えば、ある大学で、センター試験の英語の点数と、大学入学後に受けたTOEICの点数の関係性を調べたが、相関は見られなかった。センター試験の英語の点数とTOEICの点数に関係性はないのだろうか。  
⇒「その大学に入学した学生」というバイアスがかかっていたため、相関関係が見えなくなっていた。

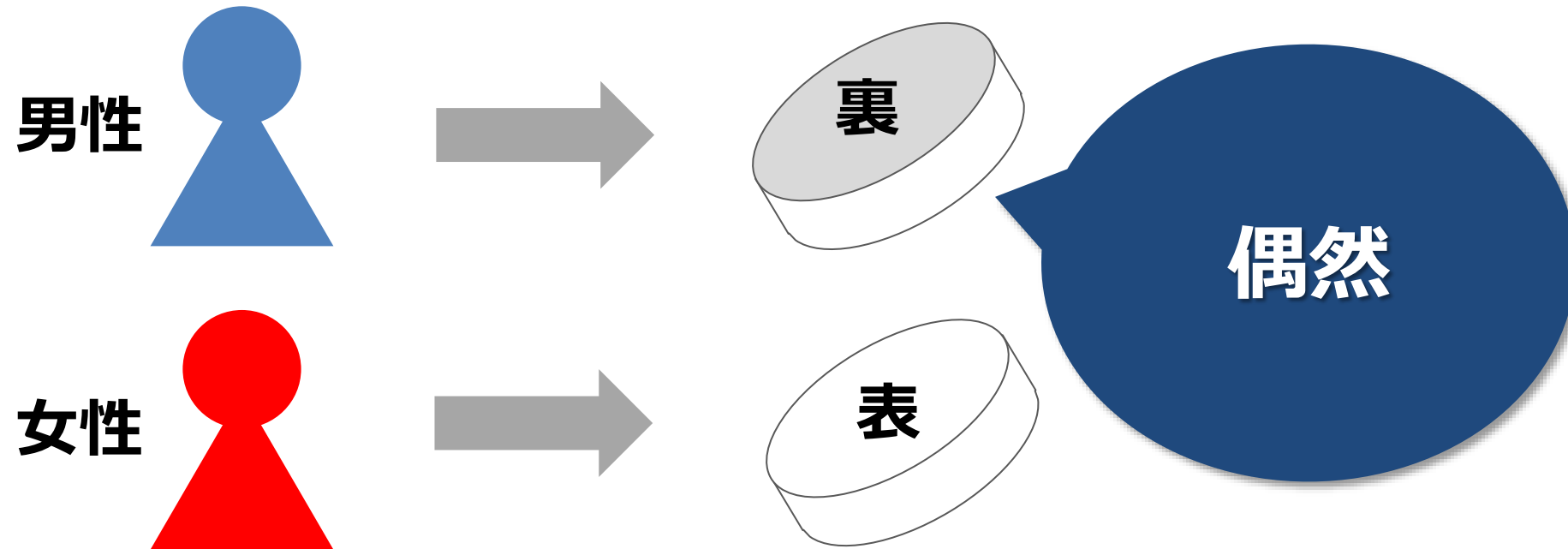


調べたい事柄とデータが一致しているか要確認

# 見せかけの相関関係が現れるケース③

## ● 偶然

- 例えば、男女がそれぞれ10回ずつコインを投げて、裏と表の出たレコードを記録していく、という作業を100回繰り返した場合、100回のうち数回は、偶然にも性別とコインの裏表に相関が見える場合がある。⇒偶然



偶然でないことを確かめることが必要

# 見せかけの相関関係が現れるケース④

## ● 因果の流れが逆

- 残業をする人とししない人とは、残業しない人の方が仕事の効率が良い。残業を禁止にすれば、仕事の効率は上がるだろうか。  
⇒「残業しないから効率が良い」のではなく「効率が良いから残業しない」



**因果関係や相関関係について考察する際には、  
背景にある共通の要因、データのバイアス、  
偶然か、因果の流れ、  
などに十分注意する必要がある。**

# 緑茶は長生きの原因か

---

- ある地域では、緑茶の消費量が多く、そこには長生きの人がたくさん住んでいる。



- 「長生きの原因は緑茶」と言えるでしょうか。
  - ここで言えるのは「緑茶を飲むことと長生きとは、何か関係がありそうだ」(＝相関関係がある)ということのみ。
  - 相関関係があるからといって、緑茶が長生きの直接的な原因であるかどうかは分からない。

**ある事象が他方の事象を引き起こしていることが  
証明できない限り、因果関係があるとは言えない**

# 問題

- おにぎりとお茶の購買履歴データがあります。
- おにぎりとお茶との間には、相関関係や因果関係はあるのでしょうか。

おにぎりを買う人はお茶も買う？

おにぎりを買えば、お茶も買うのか？

| ID | おにぎり購入 | お茶購入 |
|----|--------|------|
| 1  | 1      | 0    |
| 2  | 0      | 0    |
| 3  | 0      | 1    |
| 4  | 1      | 1    |
| 5  | 1      | 1    |
|    |        | n    |

➤ 購入を「1」、購入なしを「0」としたデータ

# 相関関係と因果関係を考察する

- 相関係数を算出したら
- クロス集計をやってみたら

$$r = 0.36$$

相関はありそう

|            | おにぎり<br>購入 | おにぎり<br>購入なし | 計    |
|------------|------------|--------------|------|
| お茶購入       | 17         | 17           | 34   |
| お茶購入<br>なし | 3          | 17           | 20   |
| 計          | 20         | 34           | 54   |
|            | 50%        | 50%          | 100% |
|            | 15%        | 85%          | 100% |
|            | 37%        | 63%          | 100% |

お茶購入者で  
おにぎり購入し  
た人と購入し  
なかった人は  
半々。

- お茶とおにぎりは関係ない？

# 相関関係と因果関係を考察する

- おにぎり購入有無で再集計してみたら…

|          | お茶購入 | お茶購入なし | 計    |
|----------|------|--------|------|
| おにぎり購入   | 17   | 3      | 20   |
|          | 85%  | 15%    | 100% |
| おにぎり購入なし | 17   | 17     | 34   |
|          | 50%  | 50%    | 100% |
| 計        | 34   | 20     | 54   |
|          | 64%  | 36%    | 100% |

おにぎり購入者の85%がお茶購入。

- 「お茶→おにぎり」の関係はないが、  
「おにぎり→お茶」の関係はありそう？

# 次回のテーマ

---

次回は

**「時系列データの見方」**

お疲れ様でした！

# 社会人のためのデータサイエンス演習

第3週：分析の具体的手法

第4回：時系列データの見方

講師名：今津 義充

# 第3週の内容紹介

## 第1回

- クロス集計の軸設定と見方

## 第2回

- 散布図と相関の調べ方

## 第3回

- 相関関係と因果関係の違い

## 第4回

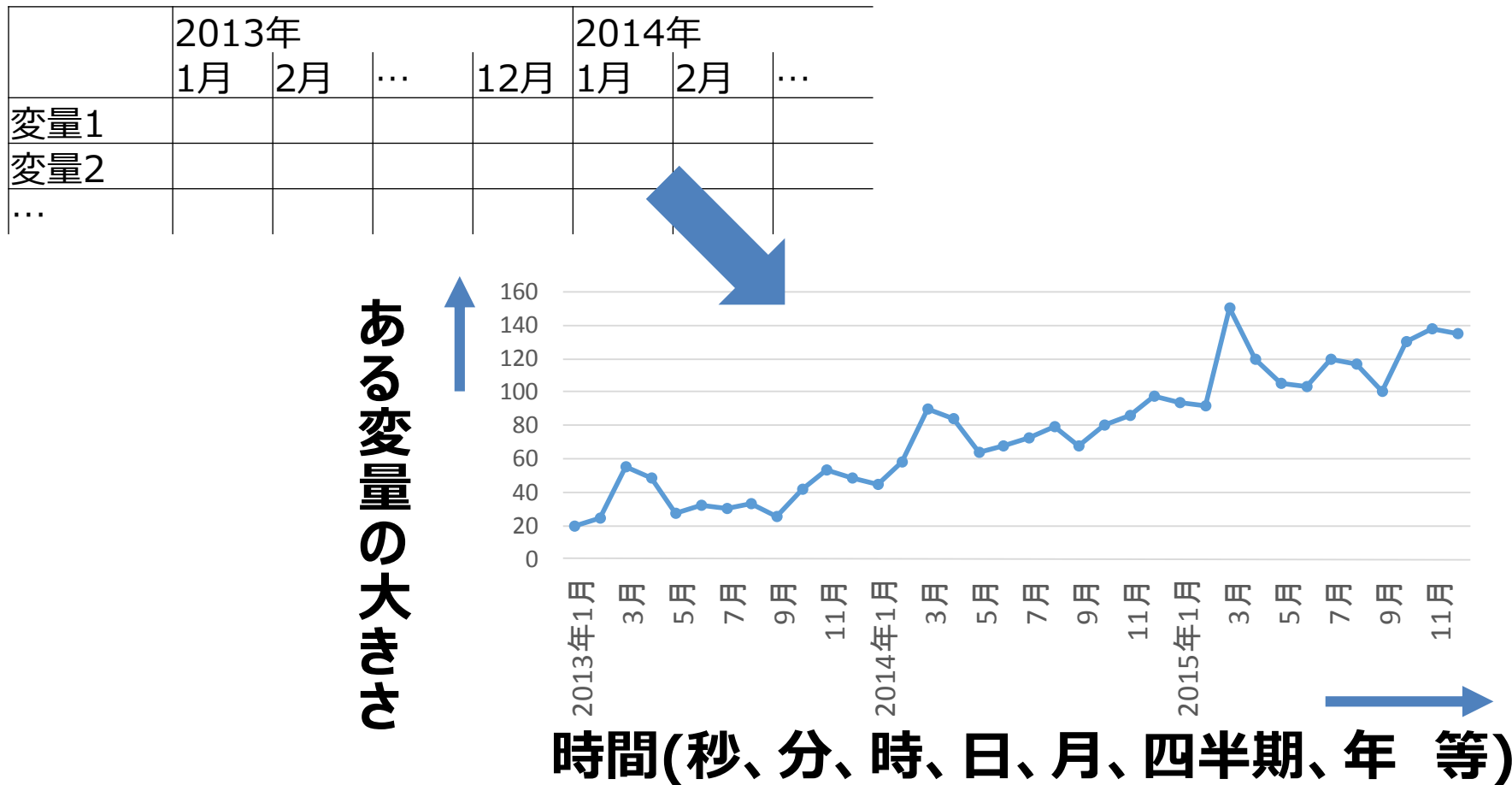
- 時系列データの見方

## 第5回

- 時系列データの分解の方法

# 時系列データとは

- 時間とともに変動する事象に対して、時間の順序で測定・観測して記録したデータを、時系列データという
  - 月別の売上、日ごとの株価、時間別のアクセス数、気温、心電図…
  - ある変量と時間の2軸で示す。



# 時系列データ分析の目的

- 時系列データ分析の目的は、過去データから変化の法則性を捉え、将来の予測を行うこと
  - 時系列データ分析は、将来予測に留まらず、過去の施策効果測定などにも利用されるケースがある(第5回：時系列データの分解の方法で説明)

精緻な時系列データ分析



有用な施策立案・実行が可能



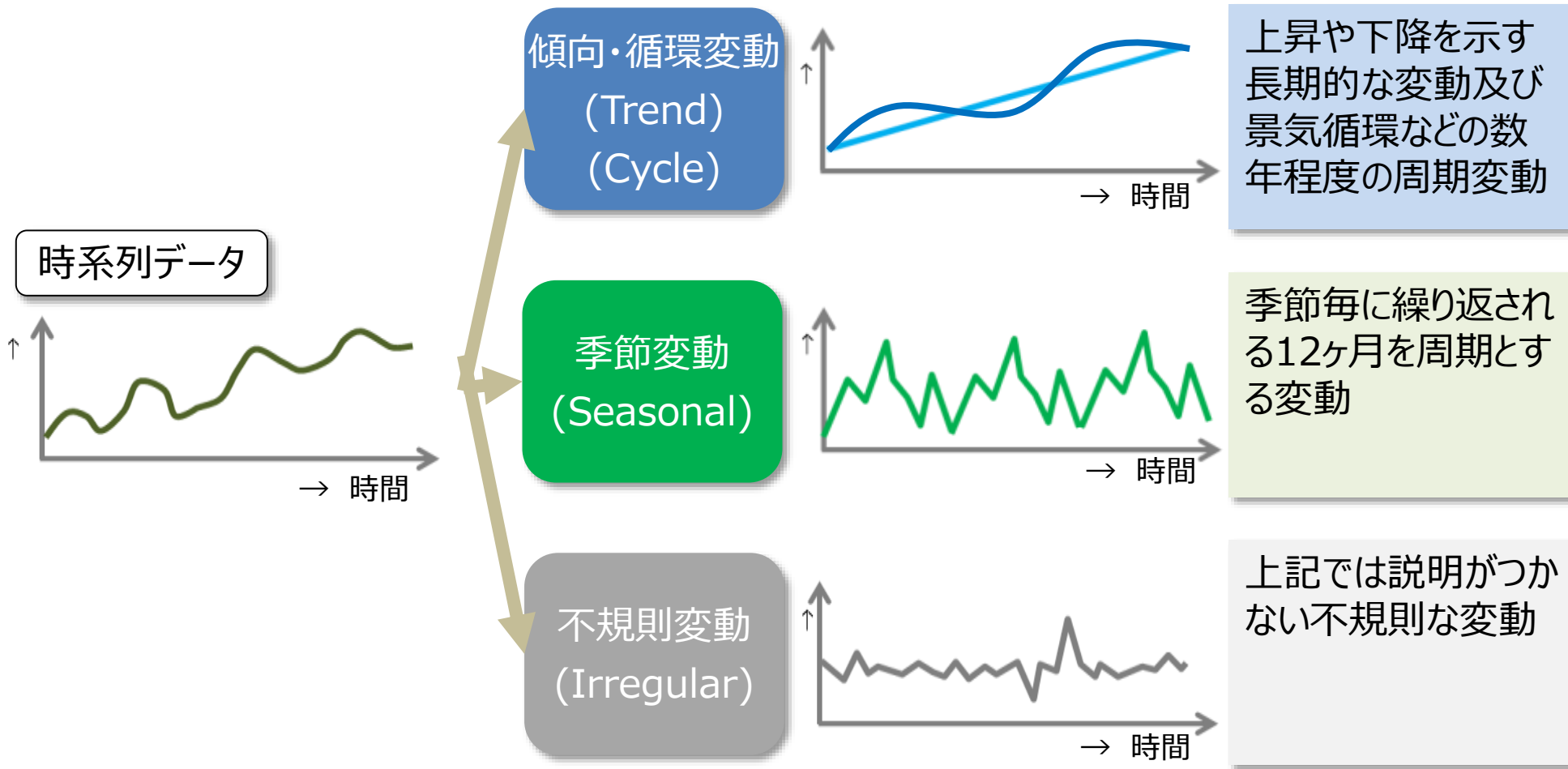
- ・的確な目標設定/KPI指標作り
- ・生産管理の見直しや組み換え
- ・効果的な広告・販促活動の促進

**未来を予測するために、過去を活用する！**

# 時系列データの構成要素とは

- 時系列データは、3つの変動要因に分解できる

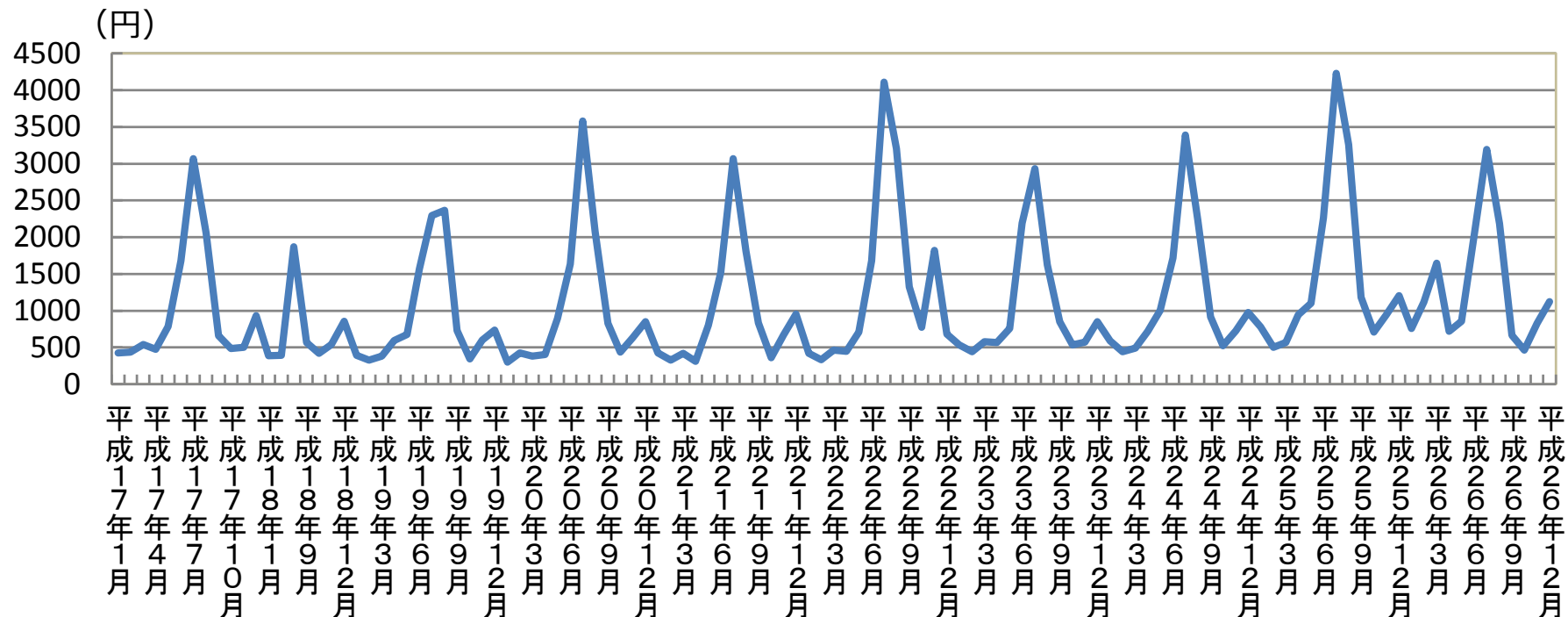
- 要素を分解することにより、過去の傾向を数式や値で表すことが可能になる。
- 過去の傾向を表した数式や値があれば、これらをもとに、将来の予測を立てることが可能になる。



# 実際に時系列データを見てみよう

- グラフは、平成17年～平成26年までの1世帯当たりの1か月間のエアコンディショナの支出額を月別に示したもの
- 周期的な傾向や推移がみられるが、読み取りが難しいので分解を実施

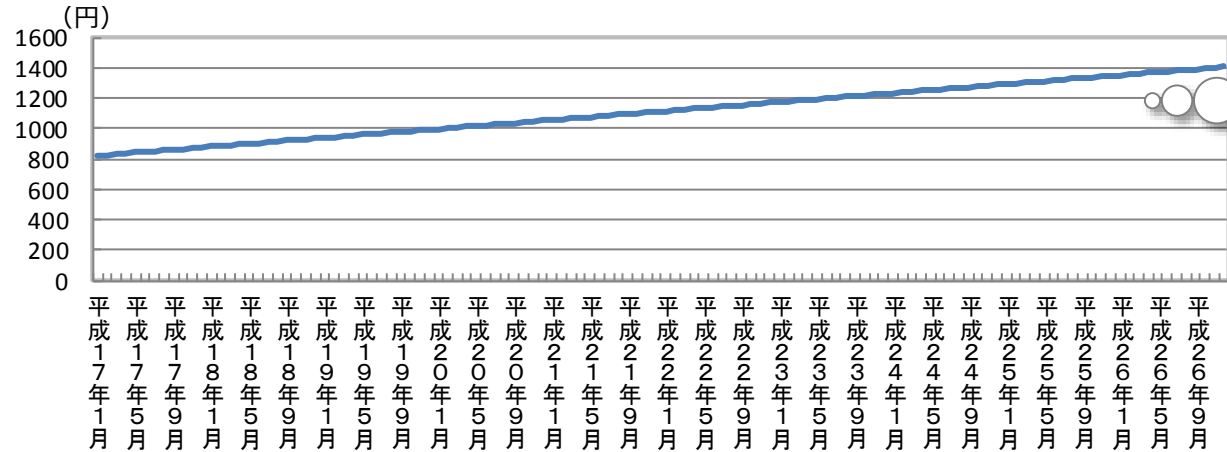
■ 1世帯当たり1か月間の支出：エアコンディショナ 全国(二人以上の世帯)



# 時系列データを分解してみると

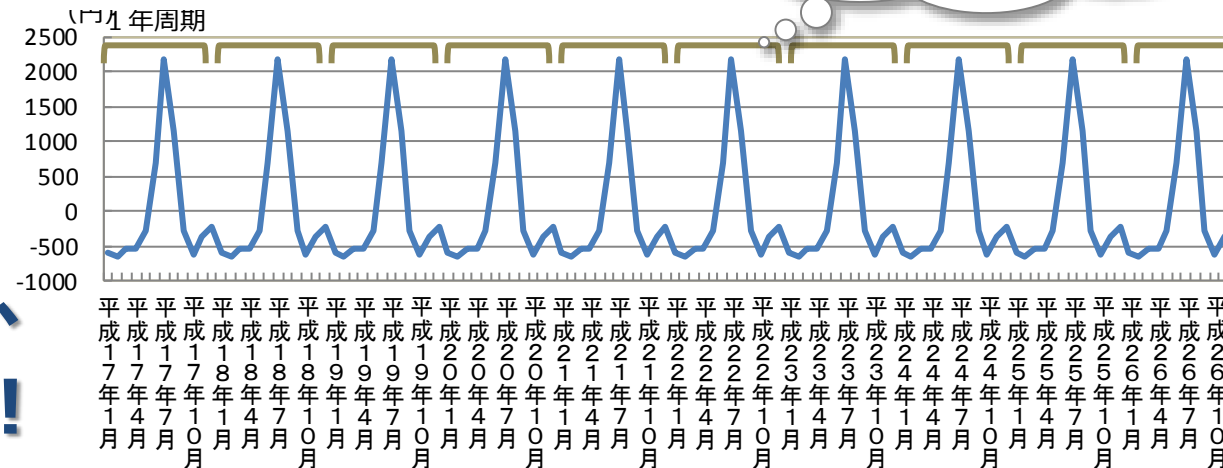
- この10年間エアコンディショナの支出額は右肩上がりで、かつ7月～8月に支出が増えていることがわかる

## ■ 傾向・循環変動



10年間でみると  
支出額は増加傾向だったのか...

## ■ 季節変動



夏は支出が増  
えていたのか...

データを分解することで、  
よりクリアに解釈できる！

# 次回のテーマ

---

次回は

**「時系列データの分解の方法」**

お疲れ様でした！

# 社会人のためのデータサイエンス演習

第3週：分析の具体的手法

第5回：時系列データの分解の方法

講師名：今津 義充

# 第3週の内容紹介

## 第1回

- クロス集計の軸設定と見方

## 第2回

- 散布図と相関の調べ方

## 第3回

- 相関関係と因果関係の違い

## 第4回

- 時系列データの見方

## 第5回

- 時系列データの分解の方法

# 各変動の算出方法①

- 時系列データを3つの変動の組み合わせとして考える

時系列データ

= 傾向・循環変動 + 季節変動 + 不規則変動  
【TC】                      【S】                      【I】

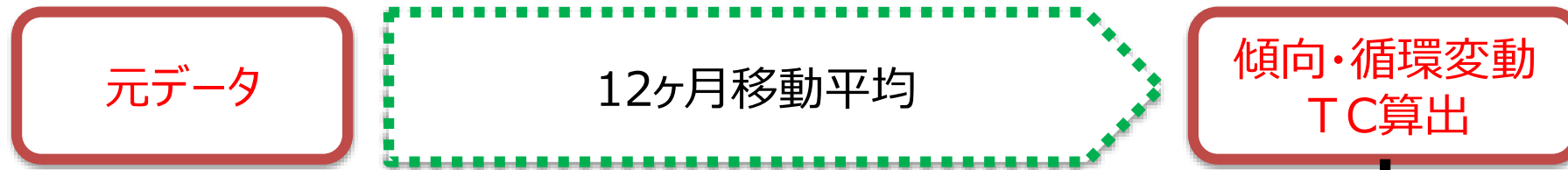
- 上記は「加法モデル」だが、この他に、時系列データを4つの変動の積と考える「乗法モデル」がある。

# 各変動の算出方法のプロセス

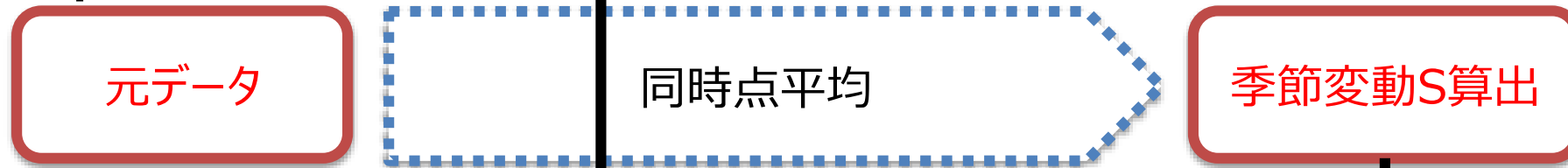
- 3つのステップで、3つの変動を求める

➤ 傾向・循環変動TC、季節変動Sを求めてから不規則変動Iを求める

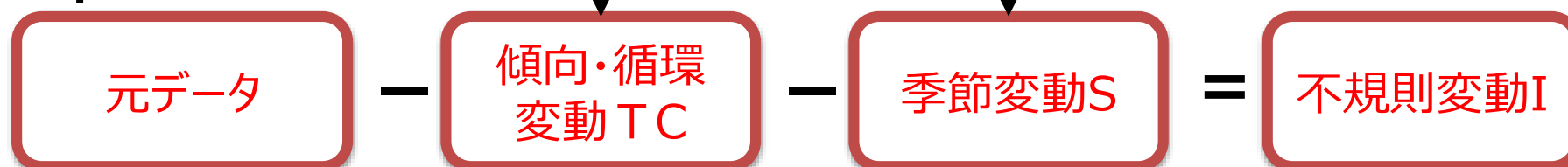
## Step1 傾向・循環変動TCの算出



## Step2 季節変動Sの算出



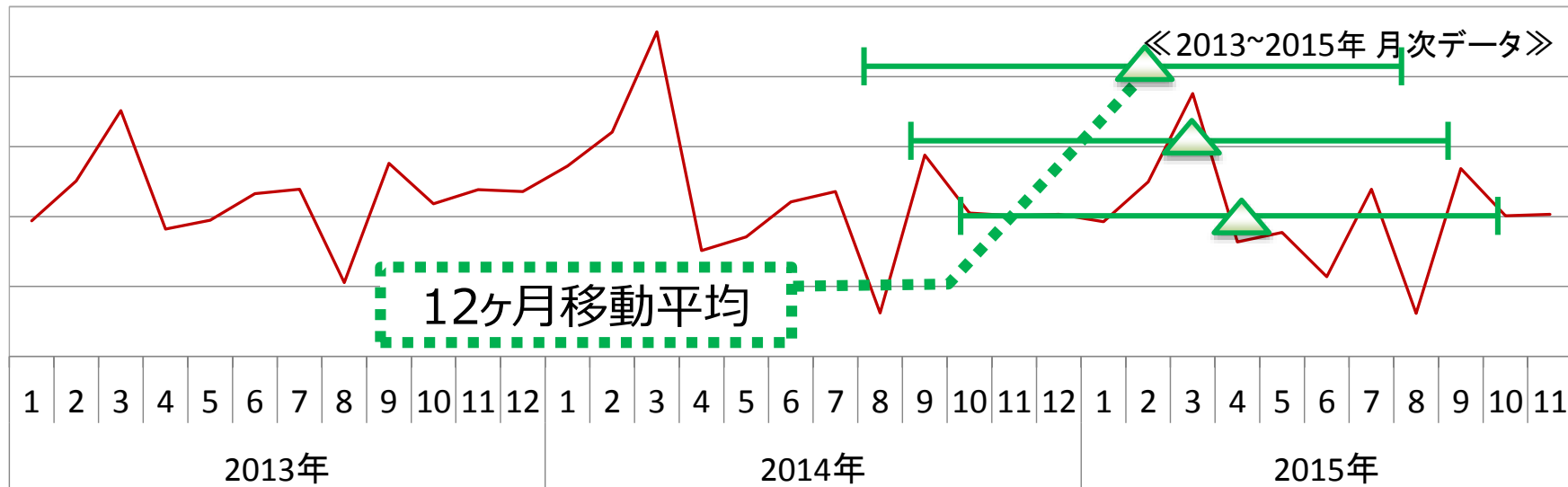
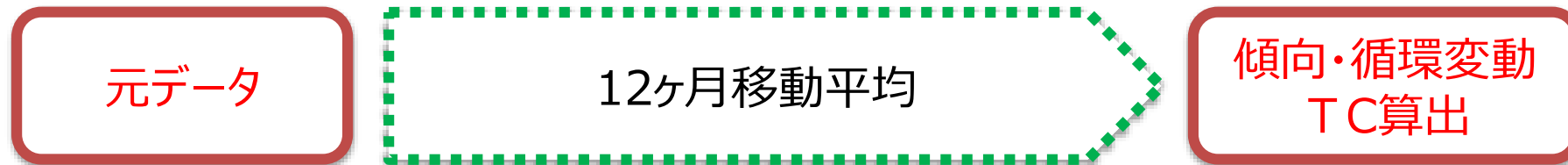
## Step3 不規則変動Iの算出



# Step 1 傾向・循環変動TCの算出方法

## ● 傾向・循環変動TCを移動平均を用いて算出する考え方

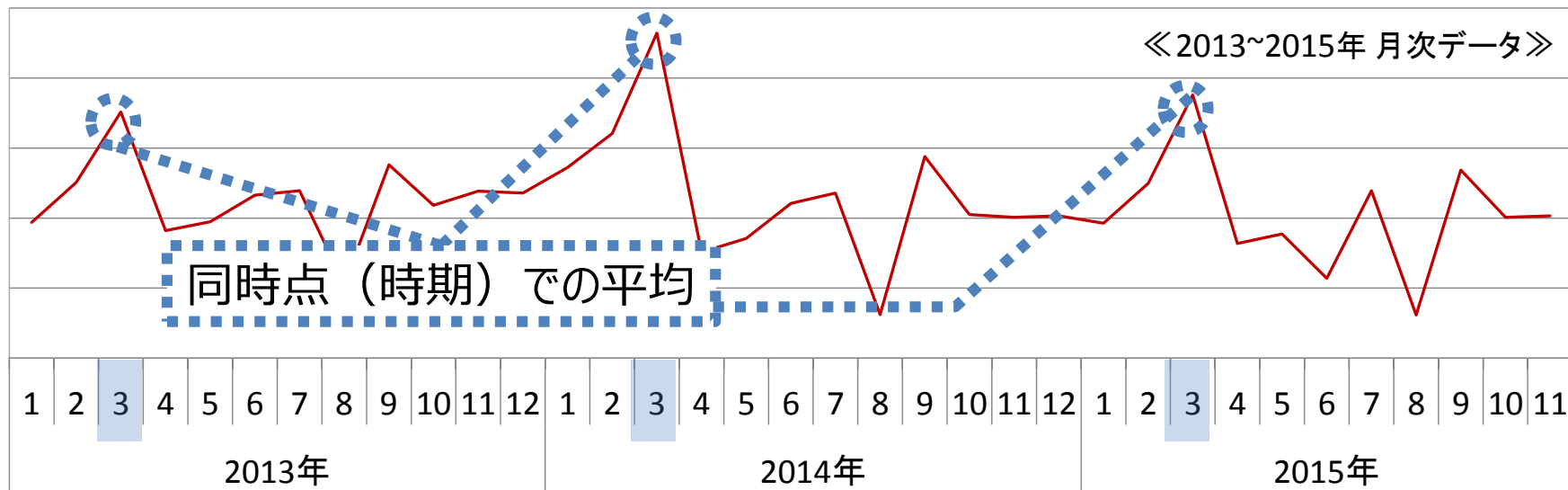
- 傾向・循環変動TCは、元データから季節変動Sと不規則変動Iを取り除けば良い
- 季節変動Sは12カ月の平均を取れば打ち消されると考える
- 不規則変動Iも、上記の平均により打ち消されると考える



# Step 2 季節変動Sの算出

## ● 季節変動Sを同時点平均を用いて算出する考え方

- 季節変動Sは、年による違いを除いた各月の値として示されるため、  
**異なる年同じ月の値の平均値**を季節変動Sとする



# Step 3 不規則変動Iの算出

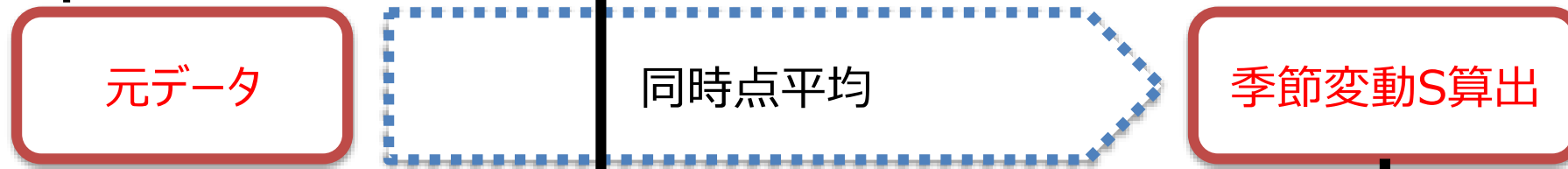
- 不規則変動は「傾向・循環変動でも季節変動でもない」もの

➤ 不規則変動を求めるには、元データから傾向・循環変動TCと季節変動Sの差を取る。

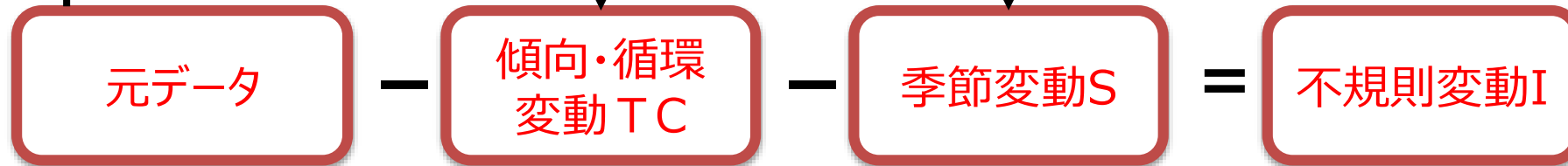
Step 1 で算出した傾向・循環変動TC



Step 2 で算出した季節変動S



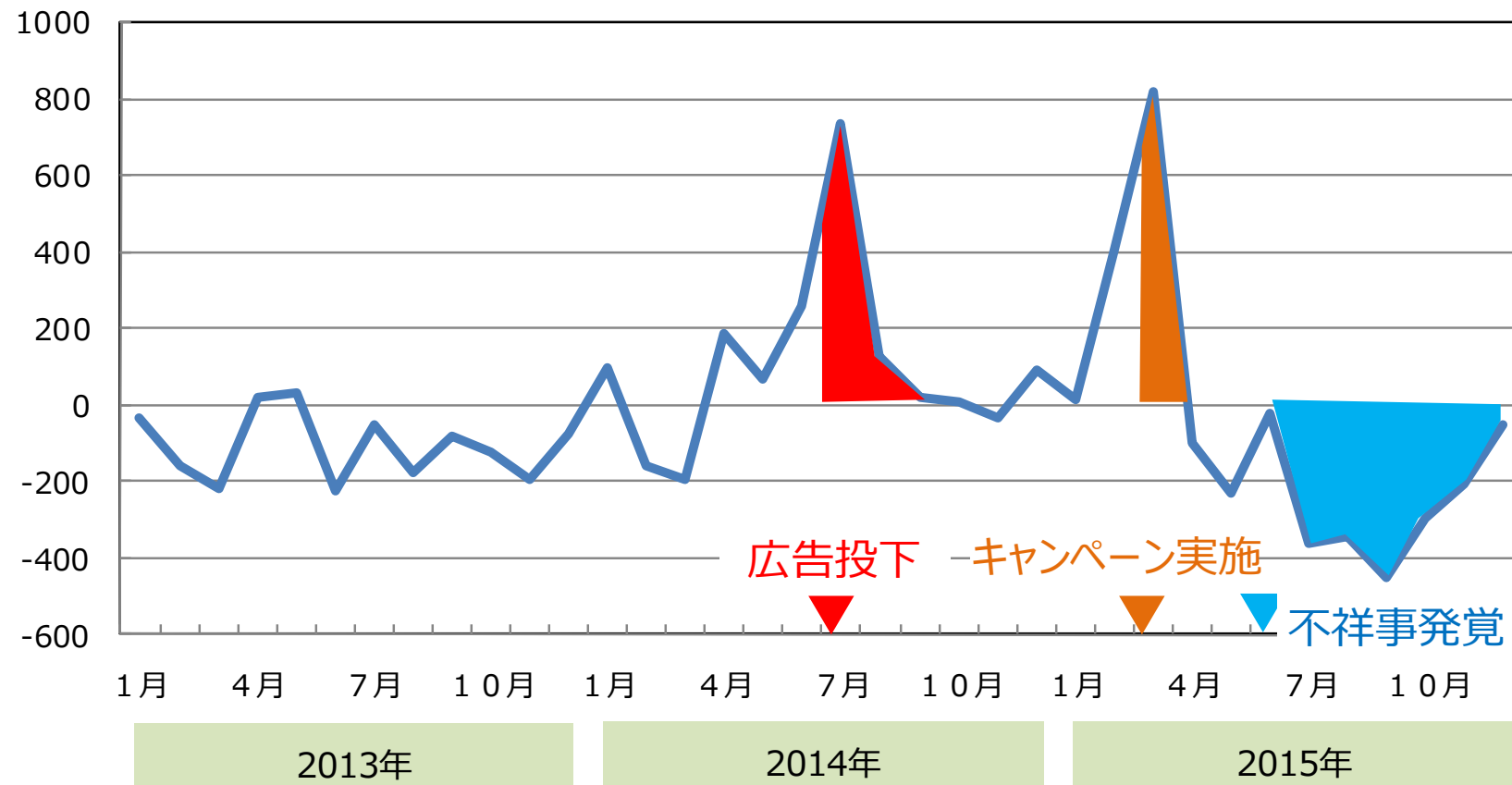
Step 3



# 不規則変動の解釈

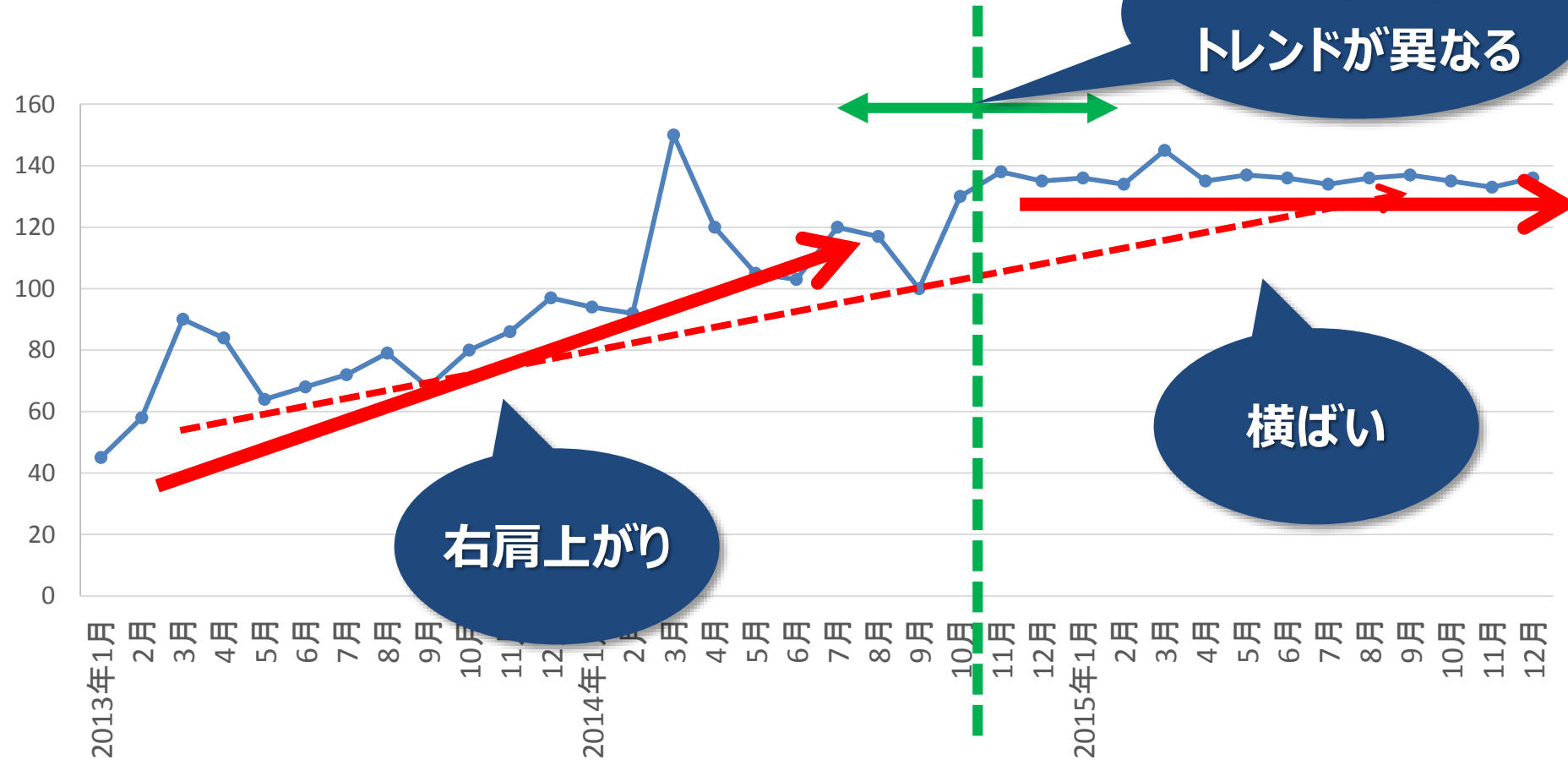
## ● イレギュラーな施策、出来事と関連付ける

- 不規則変動は、突発的出来事、自然災害、景気の短期的変動などが影響しており、説明するのは困難であるといわれている
- ただし、人為的・作為的な要因とは関連付けて分析することもできる



# 時系列データの落とし穴

## ● トrendは、右肩上がり？



時系列データを分析する際には、  
データ特性を見極めることも重要

# 次週のテーマ

---

次週は

**「ビジネスにおける予測と分析結果の報告」**

お疲れ様でした！