データリテラシー

第11回　時系列データ

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

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。

前回学んだこと

1. 相関関係
2. 相関係数
3. 因果関係と相関関係

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。

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、

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

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。

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。

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。

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。

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。

前回の問い

- 「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。

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 (ローマ字)

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

時系列データを分析できるようになる。

今回学ぶこと

1. 時系列データ
2. 移動平均
3. 階差
4. 自己相関と指数

今回の問い

- 移動平均を取ると何がわかるか？
- 階差で見れる傾向は何か？
- 自己相関とは何か？

今回学ぶこと

1. 時系列データ
2. 移動平均
3. 階差
4. 自己相関と指数

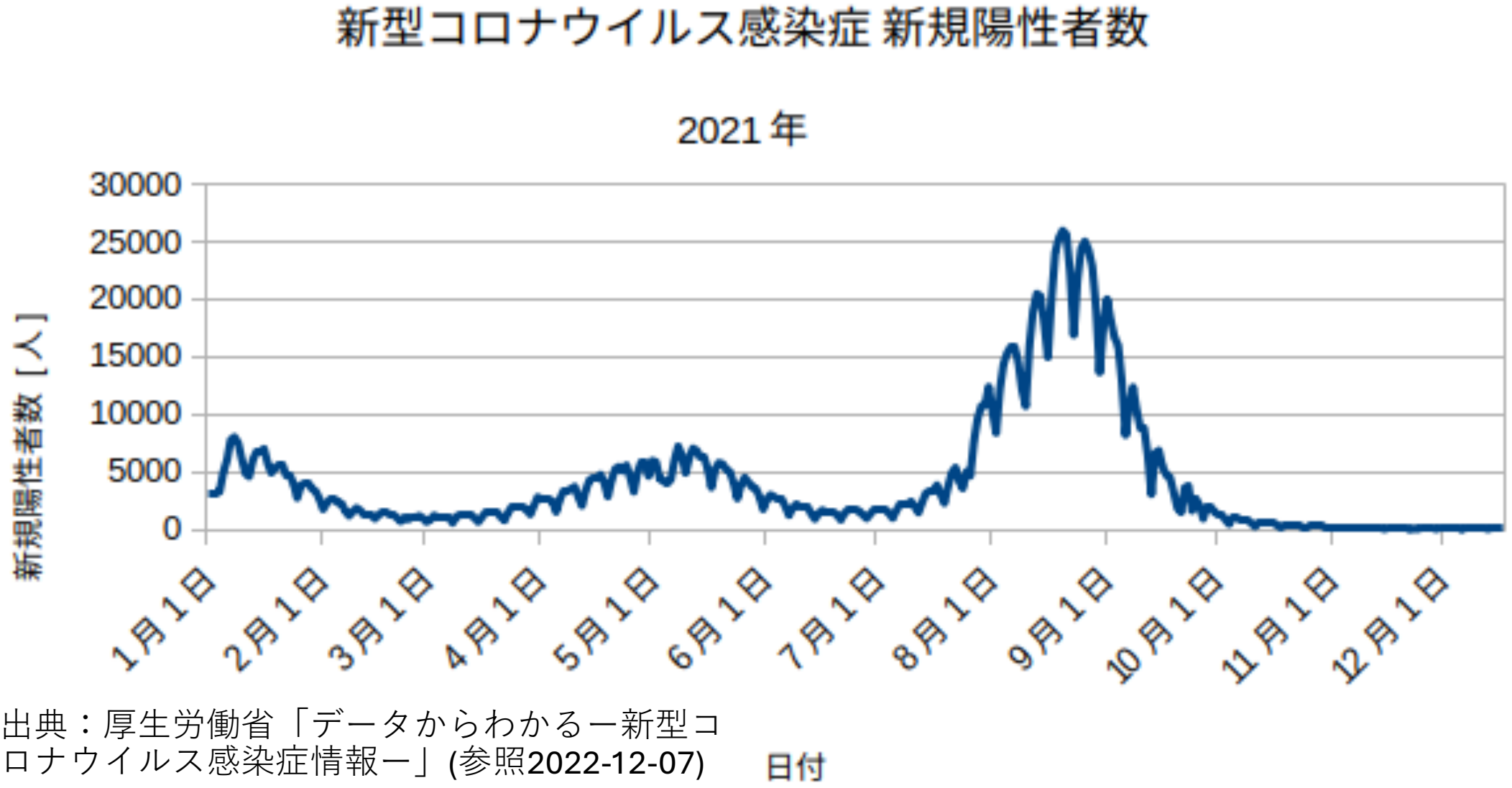
時系列データ

身の回りの現象は時間的に変動するものが多い。その現象の時間的変動を継続観測して得られるデータを**時系列データ**という。

時系列データ

時系列データは観測される時間と順序に意味がある。

時系列データの例



出典：厚生労働省「データからわかるー新型コロナウイルス感染症情報ー」(参照2022-12-07)

時系列データ

時系列データを観測された順番に$y_1, y_2, \cdots, y_n$とする。

i番目の観測値y_iとj番目の観測値y_jで$i < j$ならばy_jはy_iの後に観測されたことになる。

時系列データ

因果関係として、過去に起こったことが未来に影響を及ぼしても、未来に起こったことが過去に影響を及ぼすことはあり得ない。

つまり、$i < j$ならばy_jはy_iの結果と考えることはできても、y_iはy_jの結果とはいえない。

時系列データ

時系列データから取り出す情報として、長期的傾向と短期的傾向がある。時系列データの長期的傾向は**トレンド**という。

時系列のトレンドを見たいときは、短期的な変動は邪魔（ノイズ）になり、逆に短期的な変動を見たい時は、トレンドが邪魔になる。

演習

使用するデータセット：

https://www.kitp.org/dataliteracy/newly_confirmed_cases_daily_2021.csv

出典：厚生労働省「データからわかるー新型コロナウイルス感染症情報ー」(参照2022-12-07)

このデータは日本全体の新型コロナウイルス新規陽性者数の日別データ(2021年)である。1列目に「日付」、2列目に「日本全国の新規陽性者数」のデータが並んでいる。

ファイルの読み込み

このCSVファイルは1列目に日付があるが、ファイルで読み込む際に「日付」であることを認識させる必要がある。

ファイルの読み込み手順

1. 起動時の[**ファイルを開く**]か、メニューバーの[**ファイル**]→[**開く**]でCSVファイルを選択する。
2. 「テキストのインポート」のダイアログにおいて、「フィールド」で1列目を選択し、「列の種類」を[**日付 (YMD)**]に変更して[**OK**]ボタンをクリックする。ここで、YMDとは、「年(Year)/月 (Month)/日 (Day)」という並びのフォーマットを意味する。

フィールド

列の種類(Y): 日付 (YMD)

	日付 (YMD)	標準	標準	標準
1	Date	ALL	Hokkaido	Aomor
2	2021/1/1	3248	98	10
3	2021/1/2	3057	77	4
4	2021/1/3	3139	68	10
5	2021/1/4	3330	93	10
6	2021/1/5	4958	79	7
7	2021/1/6	6066	115	14
8	2021/1/7	7793	161	8
9	2021/1/8	8045	181	9
10	2021/1/9	7528	215	5

ファイルの読み込み手順

3. シートでは列Aに日付が表示されるはずだが、「###」と表示される場合は列幅が足りないので、列幅を広げて正しく表示されるようにする。
4. 列Aの列ヘッダで右クリックしてコンテキストメニューから[**セルの書式設定**]を選択する。
5. [**数値**]タブにおいて「カテゴリー」で[**日付**]を選択し、「形式」で[**12月31日**]を選択して[**OK**]をクリックする。

折れ線グラフ

全国の新規陽性者数の推移を折れ線グラフで表す。

データの選択

1. セルA1からB366の範囲を選択する。
2. メニューバーの[**挿入**]→[**グラフ**]を選択して「グラフウィザード」のダイアログを表示する。

折れ線グラフ

①グラフの選択

1. [**線**]を選択する。
2. 右側の4つある種類から[**線のみ**]を選択する。
3. [**次へ**]をクリックする。

折れ線グラフ

②データ範囲

はじめにデータ範囲を指定したので、ここでは何もしなくてよい。

[次へ]をクリックする。

折れ線グラフ

③データ系列

データ範囲から自動的にY値と範囲が設定されるので、ここでは何もしなくてよい。

[次へ]をクリックする。

折れ線グラフ

④グラフ要素

1. 「タイトル」に「新型コロナウイルス感染症新規陽性者数の推移」と入力する。
2. 「サブタイトル」に「2021年」と入力する。
3. 「X軸」に「日付」と入力する。
4. 「Y軸」に「新規陽性者数[人]」と入力する。
5. [**完了**]をクリックする。

これで折れ線グラフが描かれる。

目盛間隔の変更

このままだと目盛間隔が広いので、一月ごとに変更する。

1. グラフ編集モードにする。
2. グラフX軸の「1月1日」付近をダブルクリックして「X軸」のダイアログを表示する。
3. [**目盛**]タブをクリックする。
4. 「種類」を[**日付**]にする。
5. 「主間隔」の[**自動**]のチェックを外し、[**1**][**月**]にする。
6. [**OK**]をクリックする。

折れ線の属性の変更

折れ線の色などを変更してもよい。

1. グラフ編集モードにする。
2. グラフ上の青い折れ線のどこかをダブルクリックして「データ系列」のダイアログを表示する。
3. [**線**]タブをクリックする。
4. 「線の属性」の「スタイル」「色」「幅」「透明度」を変更する。
5. [**OK**]をクリックする。

今回学ぶこと

1. 時系列データ
2. **移動平均**
3. 階差
4. 自己相関と指数

移動平均

時系列データが短期的に大きくなったり小さくなったりを繰り返す振動をしている場合、各時点の前後のいくつかの観測値で平均を取ってトレンドの情報を取り出すことができる。これを**移動平均**という。

移動平均

ある時点の観測値y_iの前後$2k+1$個の観測値 (前のk個と後ろのk個)との移動平均は以下の通り。

$$z_i = \frac{y_{i-k} + y_{i-k+1} + \cdots + y_{i-1} + y_i + y_{i+1} + \cdots + y_{i+k-1} + y_{i+k}}{2k+1}$$

$$= \frac{1}{2k+1} \sum_{j=-k}^{k} y_{i+j}$$

移動平均

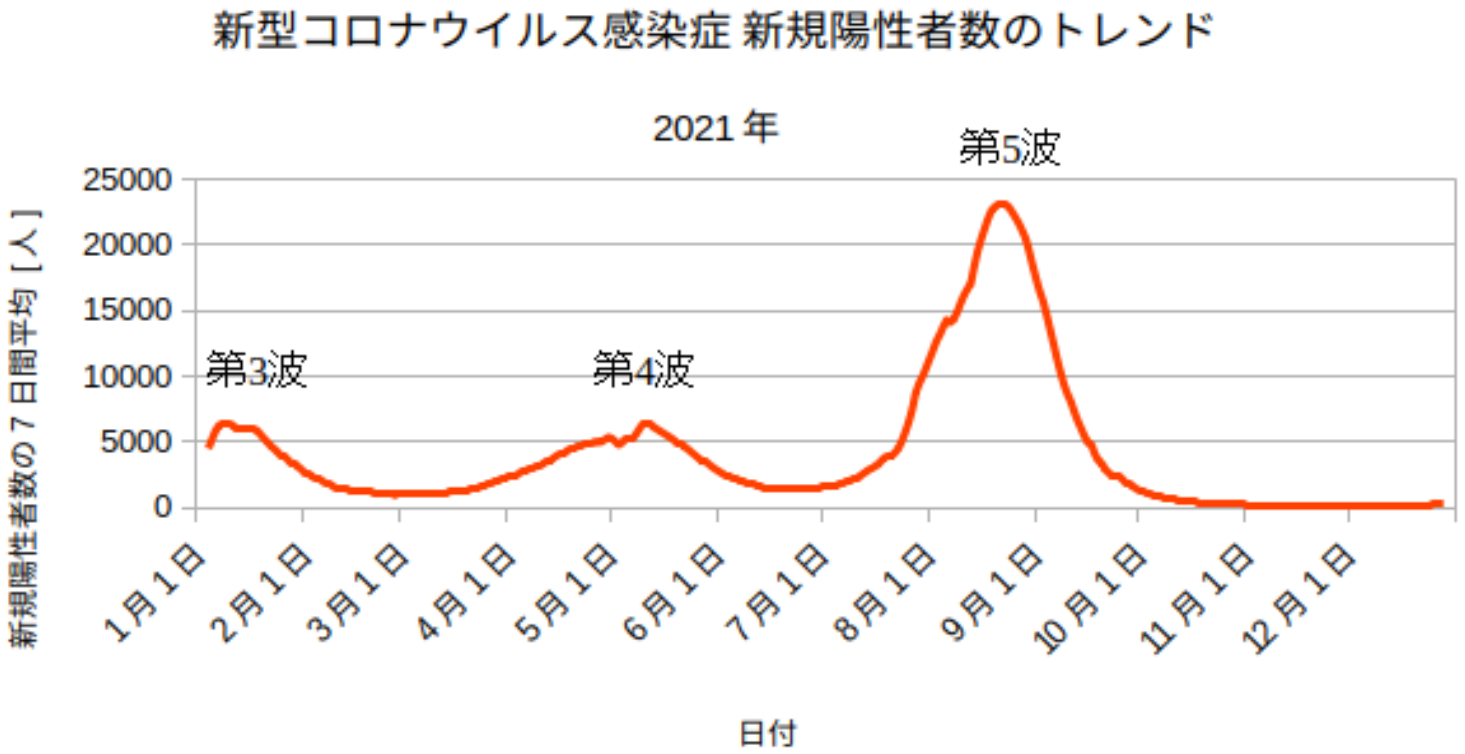
kをいくつにとれば良いかはデータによって異なるので、調整しなくてはならない。例えば$k = 3$だと、

$$z_i = \frac{y_{i-3} + y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2} + y_{i+3}}{7}$$

になる。

移動平均の例

右のグラフは日本全体の新型コロナウイルス新規陽性者数のデータから7日間平均($k = 3$)をとったグラフである。第3波、第4波、第5波の大きなピークが3つあることがわかる。



演習：7日間の移動平均

先ほどの演習データを用いて7日間の移動平均を計算する。

1. セルC1に「7日間移動平均」と入力する（ラベル）。
2. セルC5に「**=AVERAGE(B2:B8)**」と入力する。
3. セルC5をコピーしてセルC6からC363に貼り付ける（セルC364からC366には貼り付けない）。

演習：7日間の移動平均

LibreOfficeには移動平均を求める機能があるので、それを利用しても良い。

1. セルB2からB366までを選択する（入力範囲）。
2. メニューバーの[**データ**]→[**統計**]→[**移動平均**]を選択すると「移動平均」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場所のセル（例えばC1）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. 「パラメーター」は間隔に[**7**]を入力する。
6. [**OK**]をクリックすると移動平均の結果が出力される。「**#N/A**」と表示されるセルはデータ不足で計算できないところなので気にしなくてよい（削除してもよい）。

演習：7日間の移動平均（折れ線グラフ）

グラフウィザード

1. A1からC366を選択する。
2. メニューバーの[**挿入**]→[**グラフ**]を選択して「グラフウィザード」のダイアログを表示する。

演習：7日間の移動平均（折れ線グラフ）

①グラフの種類

1. [**線**]を選択する。
2. 右側の4つある種類から[**線のみ**]を選択する。
3. [**次へ**]をクリックする。

演習：7日間の移動平均（折れ線グラフ）

②データ範囲

データが選択されているので、何もしなくて良い。

1. [**次へ**]をクリックする。

演習：7日間の移動平均（折れ線グラフ）

③データ系列

データ範囲から自動的にY値と範囲が設定されるので、ここでは何もしなくてよい。

1. [**次へ**]をクリックする。

演習：7日間の移動平均（折れ線グラフ）

④グラフ要素

1. 「タイトル」に「新型コロナウイルス感染症 新規陽性者数のトレンド」と入力する。
2. 「サブタイトル」に「2021年」と入力する。
3. 「X軸」に「日付」と入力する。
4. 「Y軸」に「新規陽性者数の7日間平均 [人]」と入力する。
5. [**完了**]をクリックする。

今回の問い

- 移動平均を取ると何がわかるか？
  - トレンドの情報を取り出すことができる。
- 階差で見れる傾向は何か？
- 自己相関とは何か？

今回学ぶこと

1. 時系列データ
2. 移動平均
3. **階差**
4. 自己相関と指数

階差

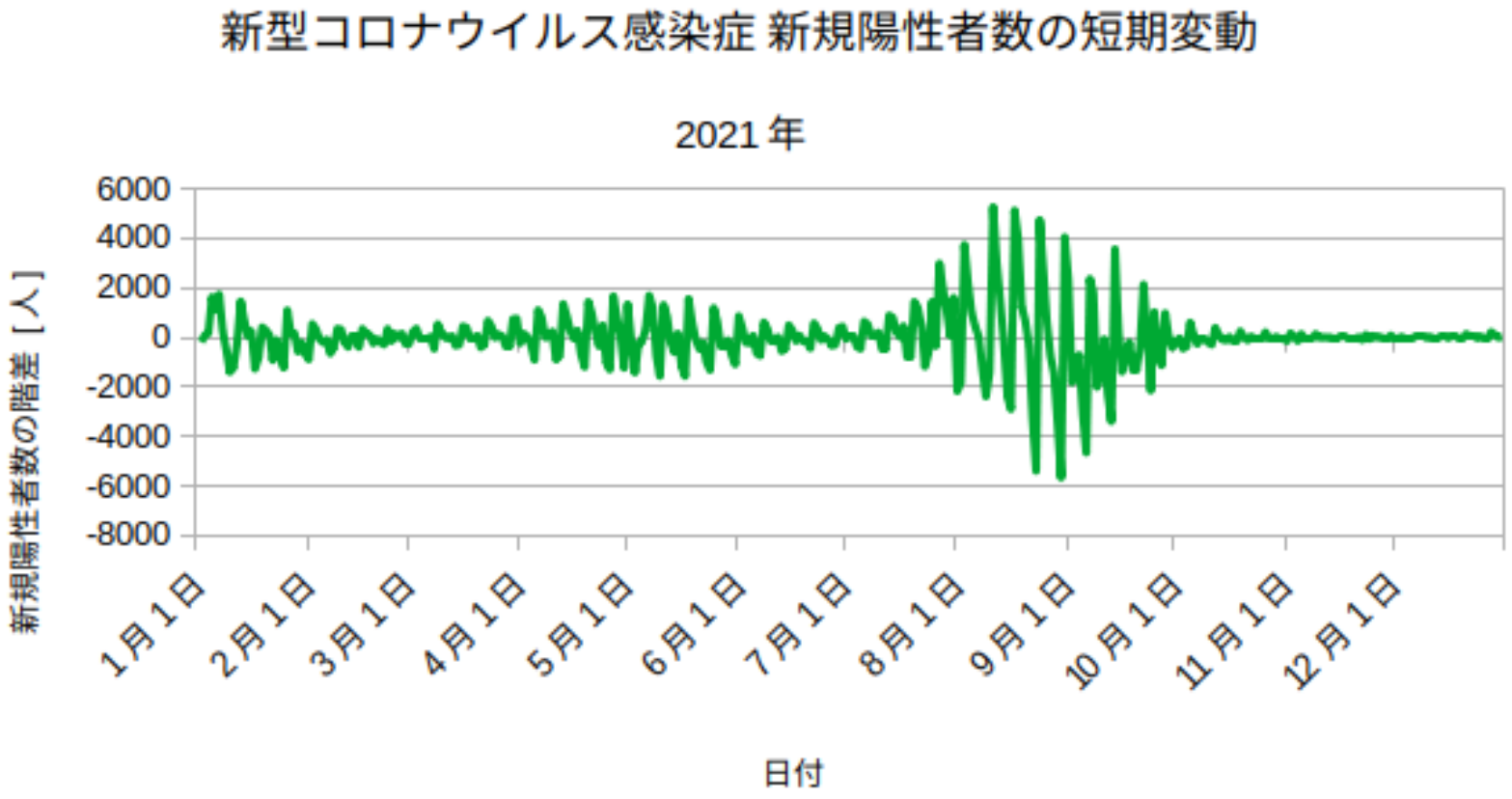
ある時点の観測値y_iとその一つ前の観測値y_{i-1}の差を**階差**といい、Δy_iで表す。

$$\Delta y_i = y_i - y_{i-1}$$

階差によってトレンドを取り除き、短期的傾向を見ることができる。

階差の例

右のグラフは日本全体の新型コロナウイルス新規陽性者数のデータから階差をとったものである。短い周期で振動しており、振動の振幅にもいくつかのピークがあることが分かる。



演習：階差

先ほどの演習データを用いて階差を計算する。

1. セルD1に「階差」と入力する（ラベル）。
2. セルD3に「**=B3-B2**」と入力する。
3. セルD3をコピーしてセルD4からD366に貼り付ける。

余裕があれば折れ線グラフを描いてみよう。

ファイルの保存

CSV形式だと数式やグラフが保存できない。ODF表計算ドキュメント（拡張子.ods）として保存しよう。

今回の問い

- 移動平均を取ると何がわかるか？
  - トレンドの情報を取り出すことができる。
- **階差で見れる傾向は何か？**
  - **トレンドを取り除き、短期的傾向を見ることができる。**
- 自己相関とは何か？

今回学ぶこと

1. 時系列データ
2. 移動平均
3. 階差
4. 自己相関と指数

自己相関

同じ時系列データの中で、異なる時点との相関を**自己相関**という。自己相関を表す指標を**自己相関係数**という。各時点とh後の時点との自己相関係数r_hは次ページのように定義される。

自己相関係数

$$r_h = \frac{\dfrac{1}{n-h}\displaystyle\sum_{i=1}^{n-h}(y_i - \bar{y})(y_{i+h} - \bar{y})}{S_y^2}$$

ここで$\bar{y}$は平均、S_y^2は分散である。

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

自己相関係数

自己相関係数は相関係数の一種なので、
相関係数の性質を持つ。

$r_h > 0$：
各時点の傾向がh後の時点でも同じ傾向

$r_h < 0$：
傾向が逆転

コレログラム

hを変化させることでr_hも変化する。横軸にh、縦軸にr_hをとって描いたグラフを**コレログラム**という。

コレログラム

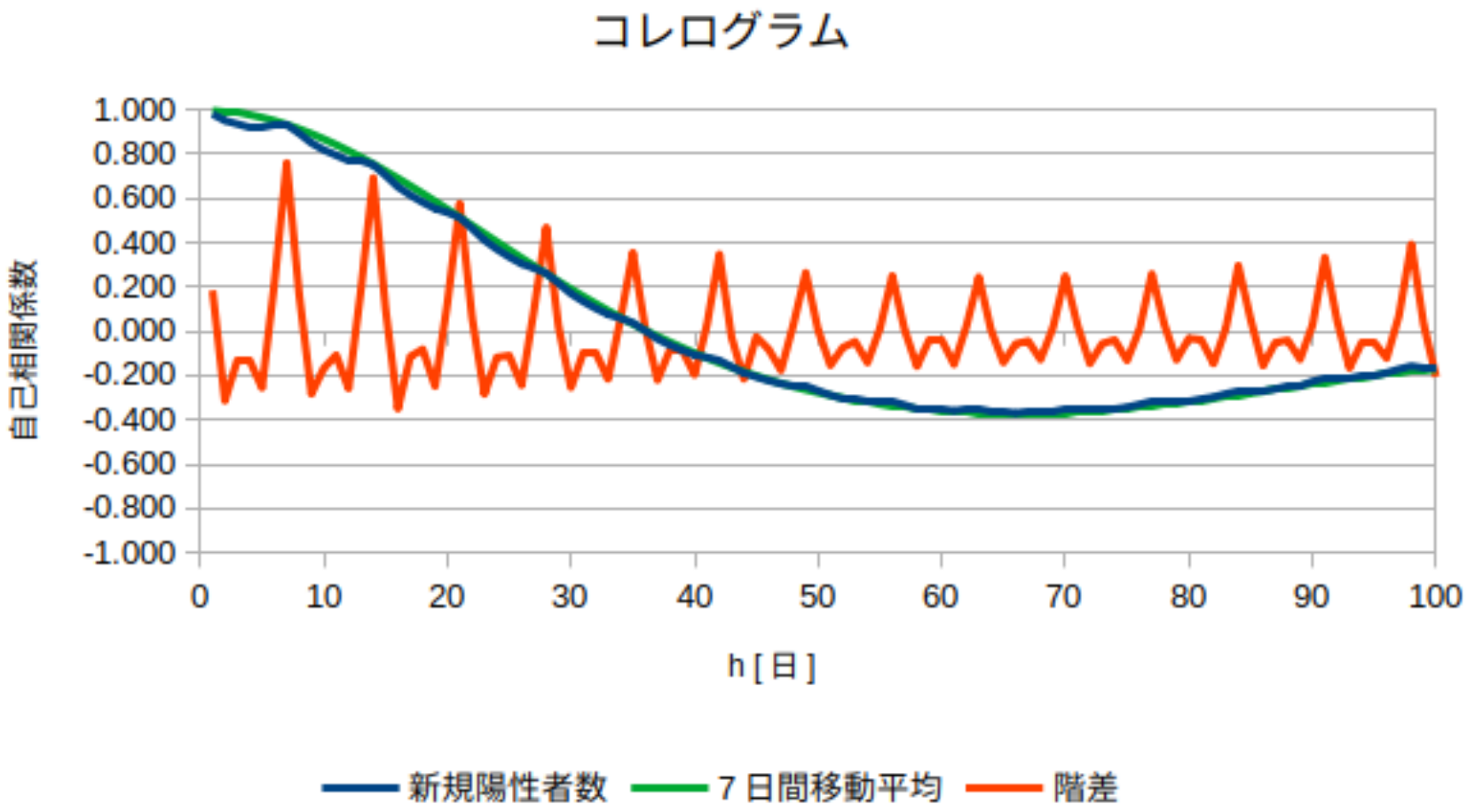
hが大きいほど、より後で観測された観測値との相関になるので、相関は弱くなるのが一般的である。

しかし、周期性を持つ時系列データでは、その周期ごとに相関は強くなる。

コレログラムを見ることで周期がどれくらいか知ることができる。

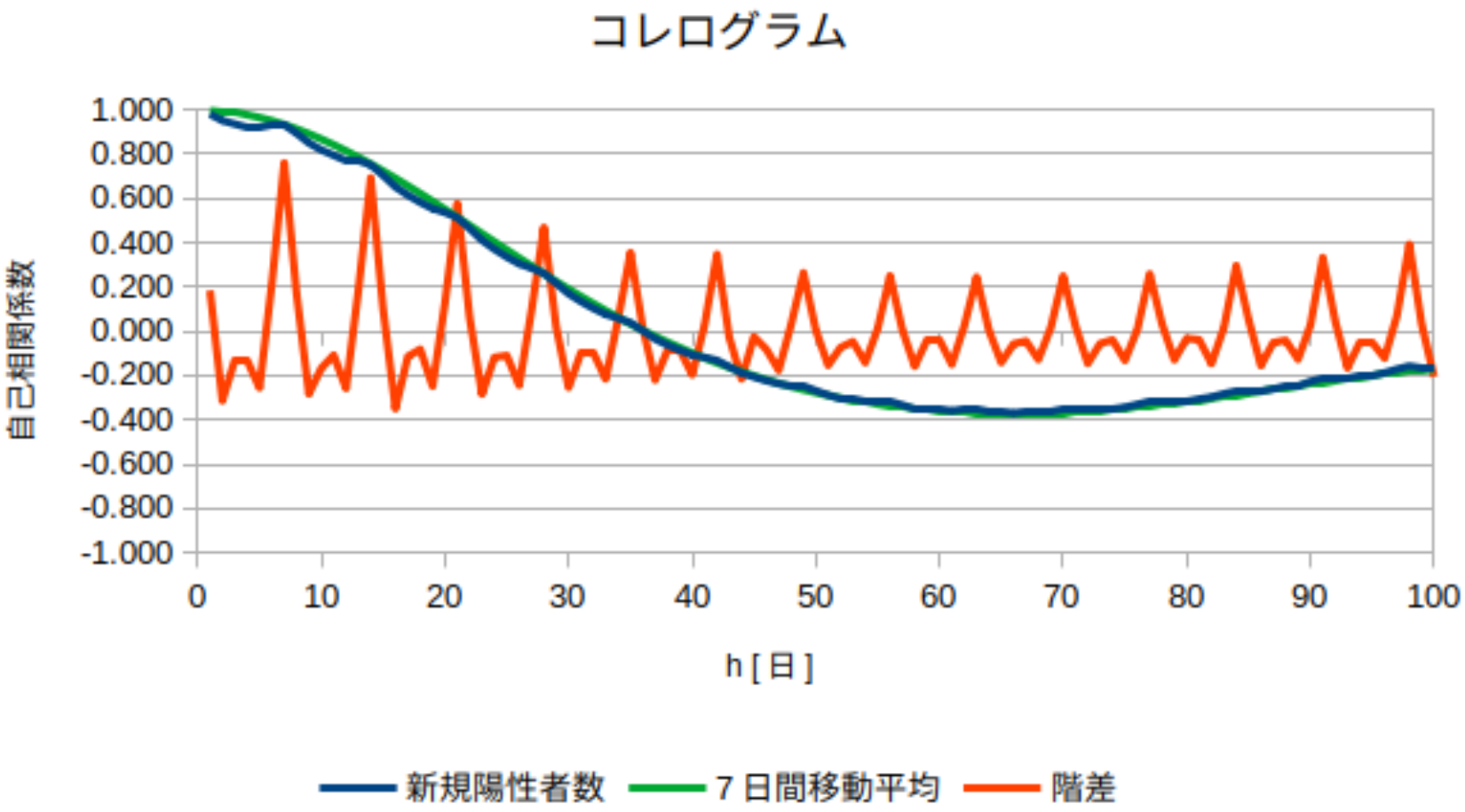
コレログラムの例

新型コロナウイルス新規陽性者数のデータのコレログラム



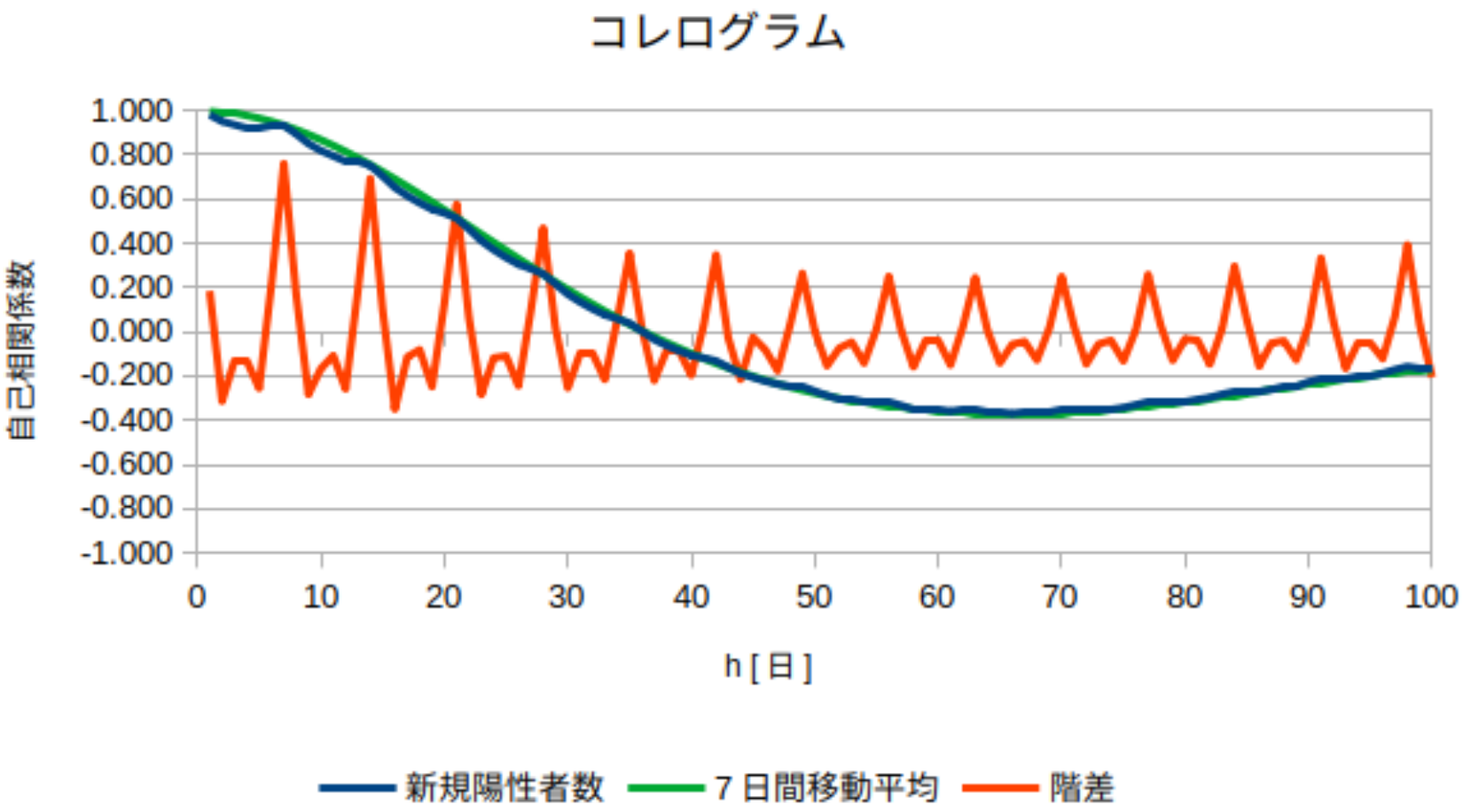
コレログラムの例

新規陽性者数とその7日間移動平均の自己相関係数にはあまり違いがなく、hが大きくなるにつれて相関は徐々に弱くなっている。



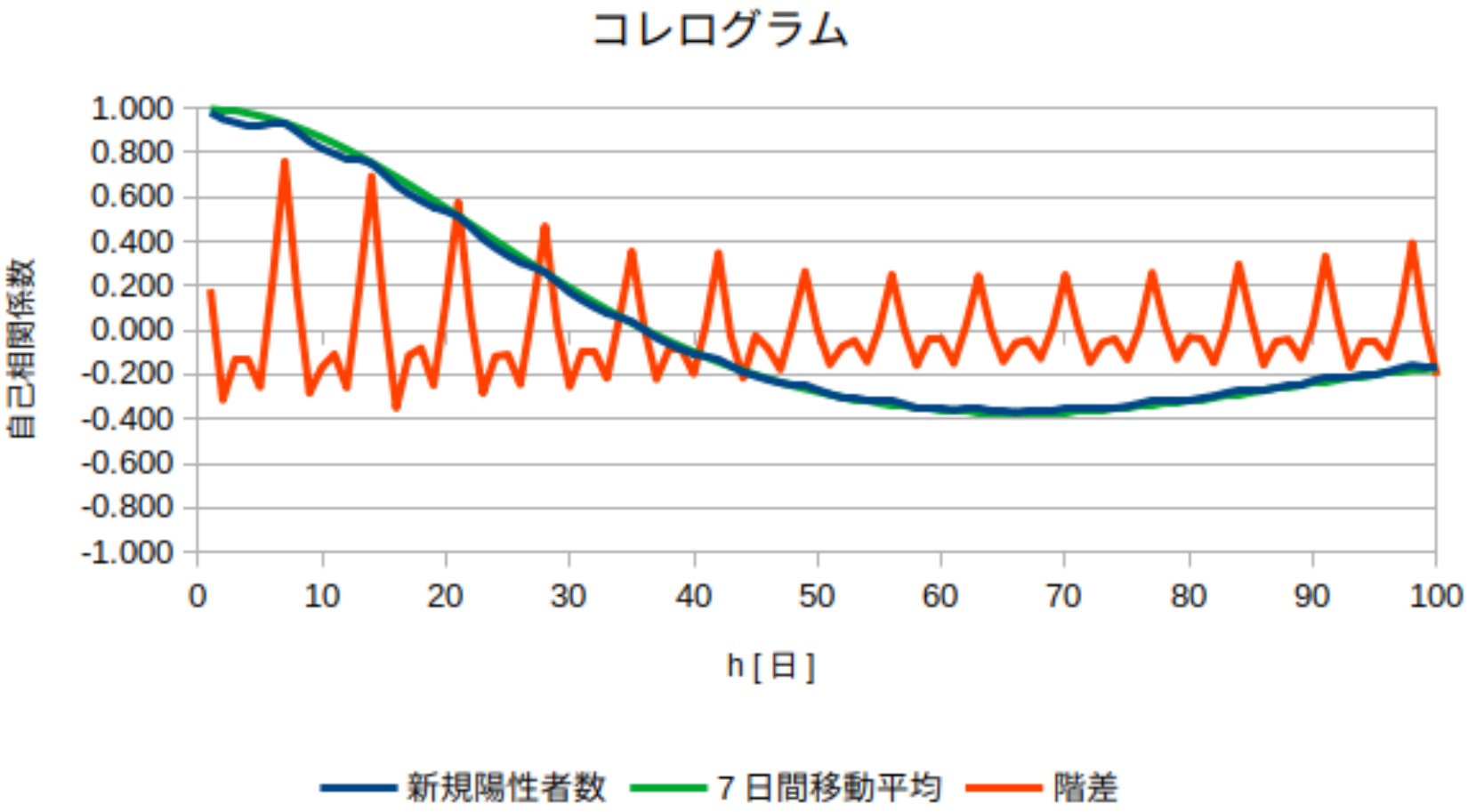
コレログラムの例

$h = 66$日にマイナスのピークがある。これは、新規陽性者数の山と谷はほぼ2ヶ月ごとに現れることを示している。



コレログラムの例

階差は$h = 7, 14, 21, 28, \cdots$と7の倍数のときに自己相関係数に鋭いピークが見られるため、短期変動の周期は7日であることが分かる。



指数

複数の時系列データを比較する際には、ある時点を基準の**100**として、そこからの変化率で見ることがある。この変化率を**指数**という。

物価指数

物価の変動を財・サービスの量と価格を元に計算される指数。消費者物価指数や企業物価指数などがある。

消費者物価指数 (CPI: Consumer Price Index)

全国の世帯が日常購入する商品やサービスの物価変動を表す指数。

ある基準となる時点（基準時）に家計で購入した商品やサービスと同じものと、比較する時点（比較時）に買い揃えるとしたらどれだけお金がかかるかという考えで計算されている（ラスパイレス方式）。

消費者物価指数 (CPI: Consumer Price Index)

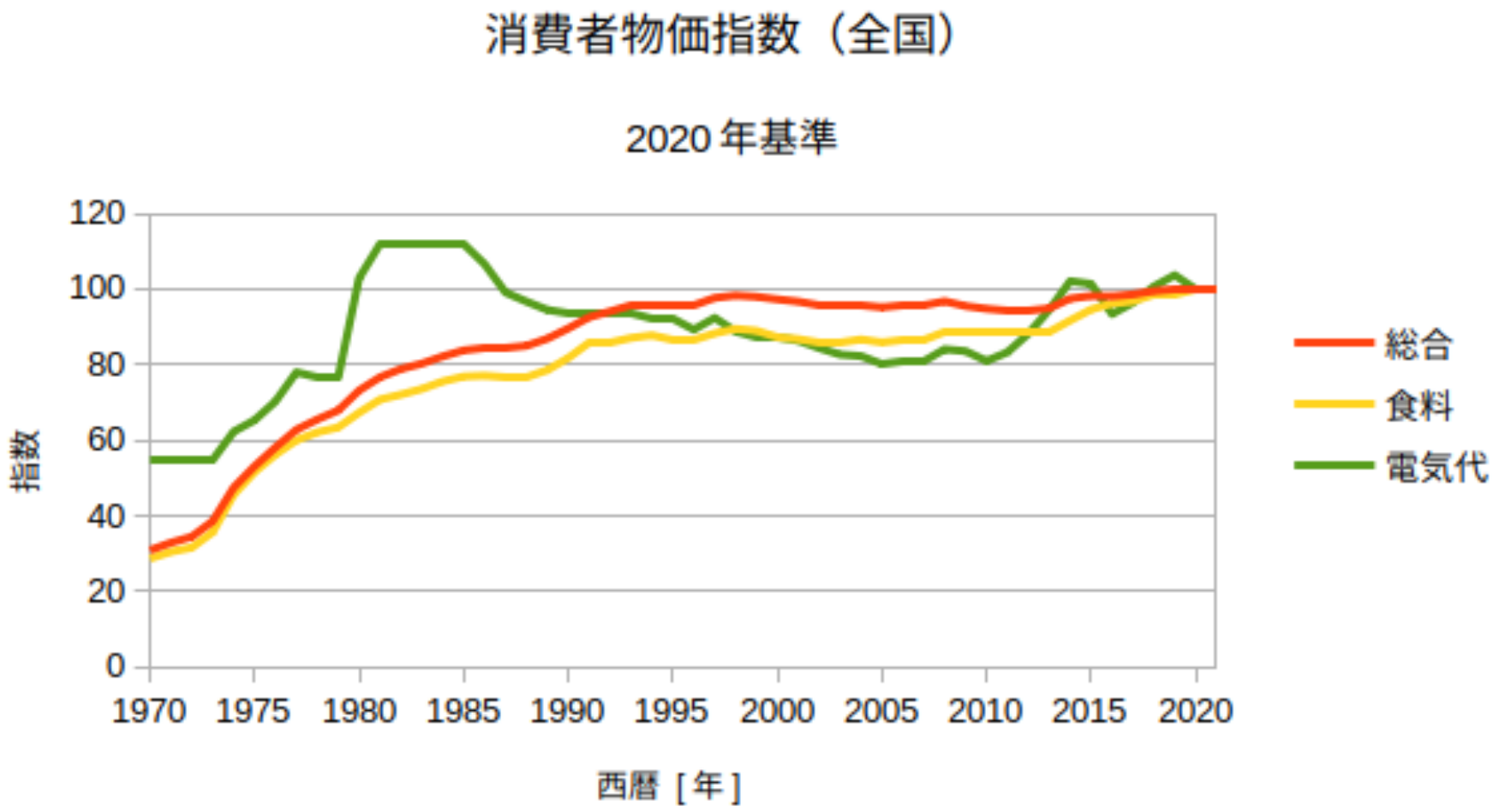
比較時の消費者物価指数は、基準時の物価を100として表される。ラスパイレス方式の計算式は以下のようになる。

$$I_t = \frac{\sum_{i=1}^{n} p_{t,i} q_{0,i}}{\sum_{i=1}^{n} p_{0,i} q_{0,i}} \times 100$$

変数	意味
I_t	比較時の消費者物価指数
i	品目
$p_{t,i}$	比較時の品目の価格
$p_{0,i}$	基準時の品目の価格
$q_{0,i}$	基準時の品目の購入数量

消費者物価指数の例

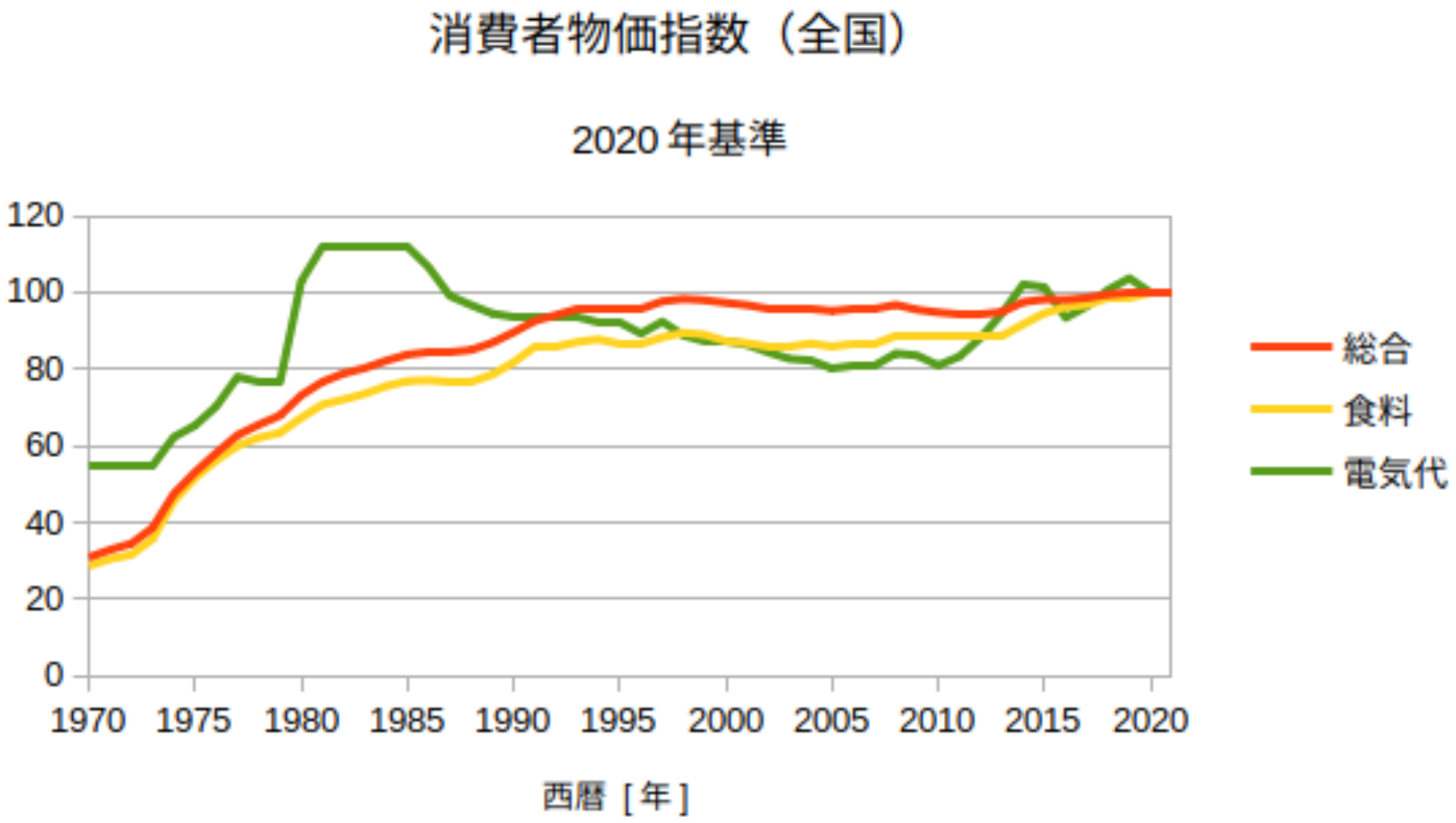
2020年を基準(100)とした消費者物価指数の推移



出典：統計局「消費者物価指数(CPI)」(参照：2022-12-03)

消費者物価指数の例

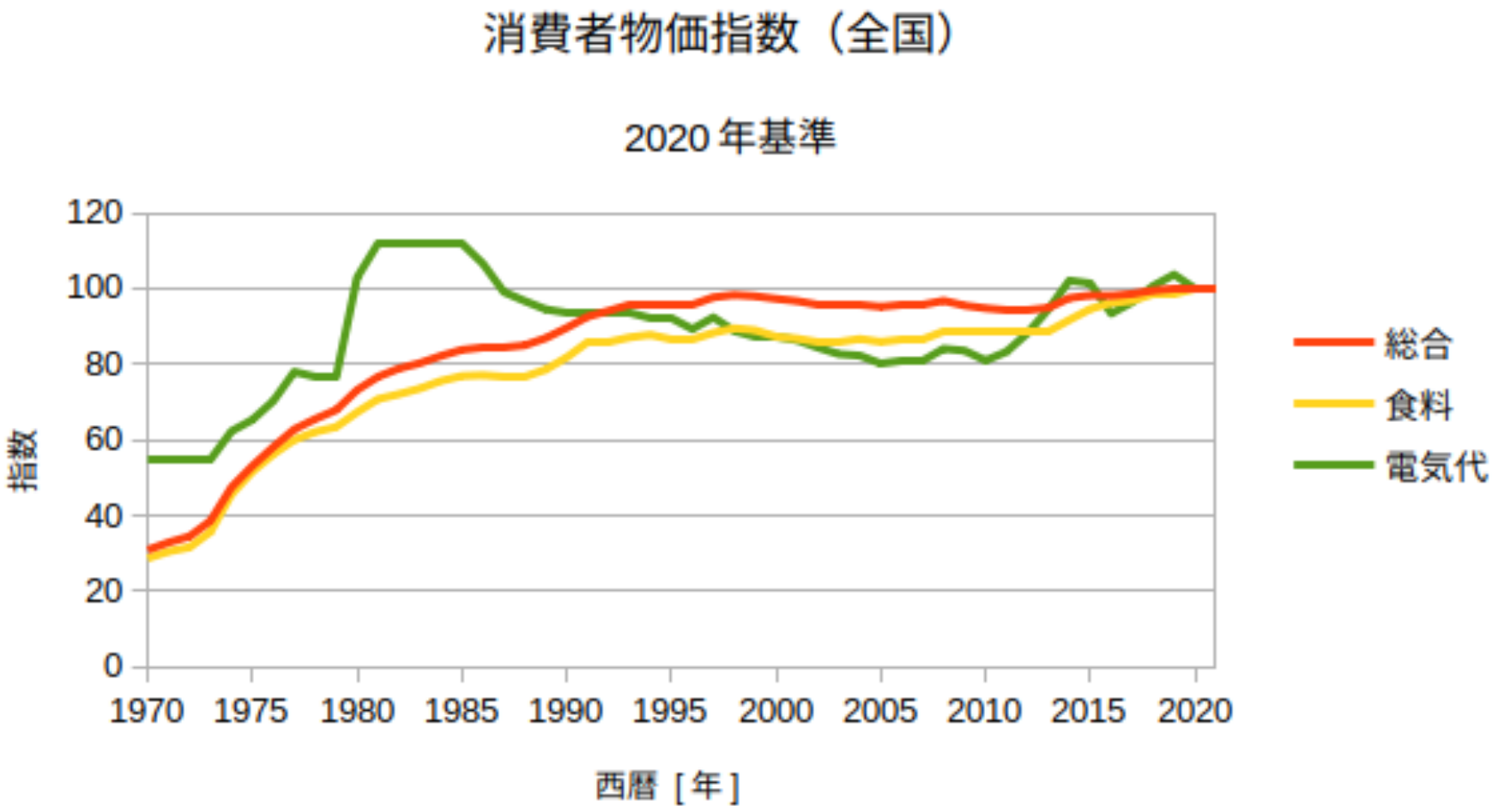
電気代が1980年に急激に上がっているのは第2次オイルショックによる。



出典：統計局「消費者物価指数(CPI)」(参照：2022-12-03)

消費者物価指数の例

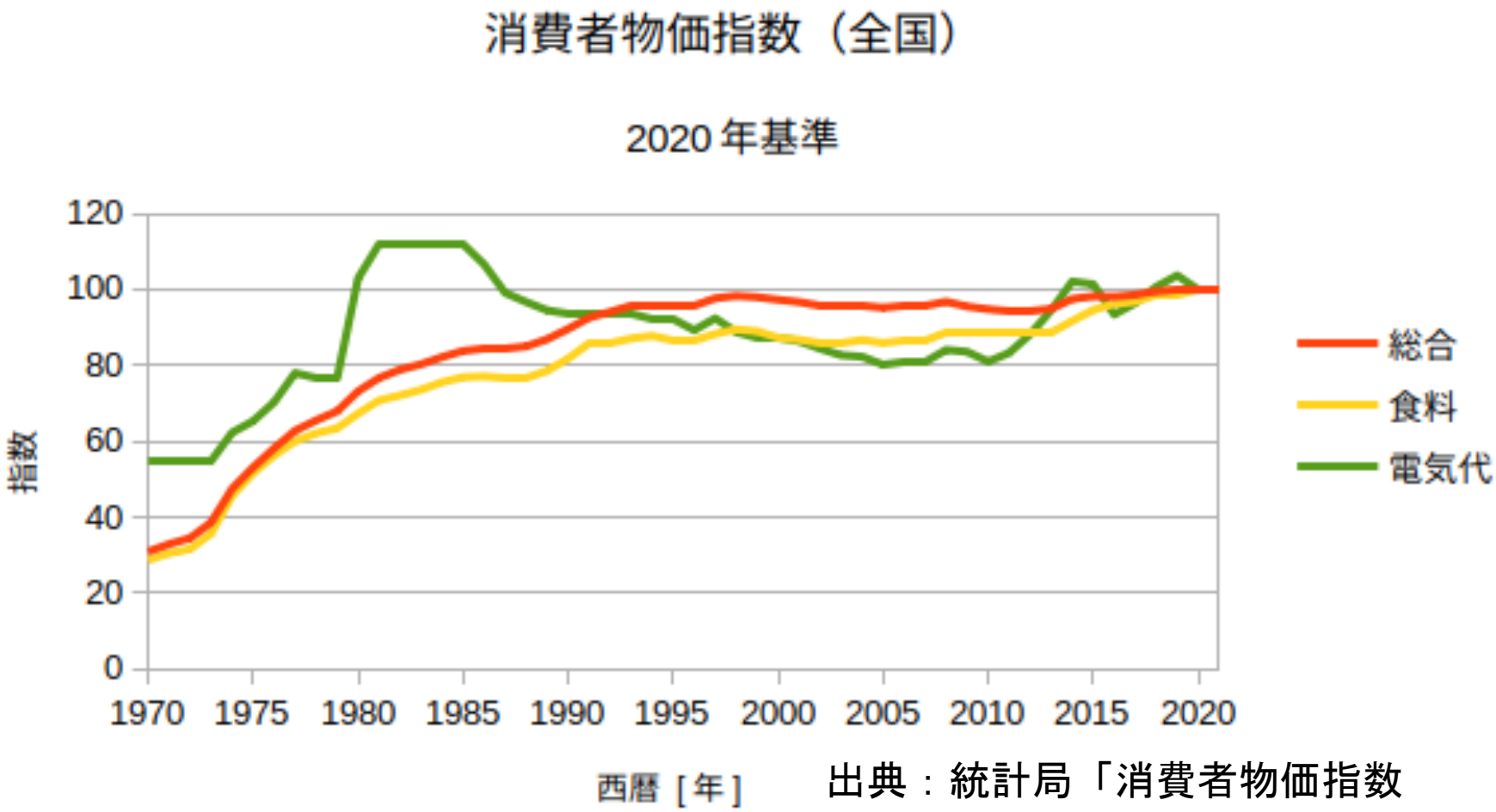
当時日本の発電は石油火力発電の割合が大きかったため、原油価格の高騰によって電気代が上がった。



出典：統計局「消費者物価指数(CPI)」(参照：2022-12-03)

消費者物価指数の例

その後、石油への依存割合を減らして天然ガスや石炭の火力発電、原子力発電などの電源を多様化することで電気代は下がった。



出典：統計局「消費者物価指数(CPI)」(参照：2022-12-03)

今回の問い

- 移動平均を取ると何がわかるか？
  - トレンドの情報を取り出すことができる。
- 階差で見れる傾向は何か？
  - トレンドを取り除き、短期的傾向を見ることができる。
- **自己相関とは何か？**
  - **同じ時系列データの中で、異なる時点との相関。**

今回学んだこと

1. 時系列データ
2. 移動平均
3. 階差
4. 自己相関と指数

今回の目標

時系列データを分析できるようになる。

課題

新型コロナウイルスの新規陽性者数のデータから「2022年」「山形県」の「新規陽性者」「7日間の移動平均」「階差」の3つの折れ線グラフを描きなさい。

課題

新型コロナウイルスの新規陽性者数のデータは以下よりダウンロードする。

https://covid19.mhlw.go.jp/extensions/public/index.html

1. このページの「新規陽性者数の推移（日別）」にある「オープンデータ」をクリックしてCSVファイルを保存する。
2. 列Aに「Date (日付)」があるので、2022/1/1から2022/12/31までを用いること。
3. 列Hに「Yamagata (山形県)」の「新規陽性者数」のデータが並んでいる。

課題

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11回課題 [学籍番号]

締め切り：12月25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