第 2 章
表計算ソフトウエア：地域別・時系列データの分析

2026 年 11 月 2 日

学習目標

(1) 平均値，中央値，標準偏差，Z 値 (偏差値) を計算してみる．
(2) 度数分布表，ヒストグラムを作成する．
(3) 2 系列の散布図，相関係数を計算し，相関関係と因果関係を理解する．
(4) INDEX 関数など範囲の処理を理解し，相関係数行列（相関図）を作成する．
(5) ソート・順位を計算できるようになる．
(6) 単回帰分析を体験する．
(7) クラスター分析の分析手順を理解する．
(8) 時系列データを分析してみる．

本章は，専修大学商学部の高萩栄一郎の著作である．

1　家計調査データから消費支出額と 5 品目を選択

家計調査から 2021 年の都道府県県庁所在地，政令指定都市別の 2 人以上の世帯の品目別の消費支出金額を分析します．そのデータの中から，家計の消費支出額と 5 品目の消費支出額の系列を選択し，分析をします．

このファイルは教科書のページからダウンロードできます (S110.xlsx).

	A	B	C	D	E	F	G
1		1	2	3	4	5	6
2	地域区分	消費支出	米	パン	ケーキ	せんべい	チョコレート
3	全国	3348287	21862	31353	7716	5719	6664
4	札幌市	3220749	27371	27609	8339	4427	6731
5	青森市	2941401	20934	25662	6563	5955	4861
6	盛岡市	3274937	22927	28638	7984	6249	5678
7	仙台市	3410054	21594	30170	8200	6141	6386
8	秋田市	2973311	21126	23775	6337	4623	5837
9	山形市	3856930	24447	26818	8064	7094	6946
10	福島市	3512916	22450	26215	7573	7112	7461

図 1　分析用データ (ワークシート名:dataset)

S110.xlsx のワークシート「dataset」に，ワークシート「全品目」から消費支出額と適当な 5 品目の列をコピーして，図 1 のようなワークシートを作成し，分析を行います．5 品目は例題の品目と異なっても同様に分析可能です（練習問題で，他の品目を学修することがあります）．本例題では，下記 5 品目を分析します．なお，1 行目の B1:G1 の 1 から 6 までの数値は，何列目になるのかを示すために，表示したものです．例題では，消費支出 (B 列)，米 (E 列)，パン (F 列)，ケーキ (FZ

列)，せんべい (GD 列)，チョコレート (GH 列) の系列の値を利用しています．

(1) ワークシート「全品目」の「消費支出」の列，B1:B54 をコピーし，ワークシート「dataset」の B2:B55 に貼り付け
(2)「米」の場合，ワークシート「全品目」の「米」の列，E1:E54 をコピーし，ワークシート「dataset」の C2:C55 に貼り付け
(3)「dataset」D,E,F,G 列も同様に作成．

動画: 分析用データの複写（音声なし）

e-Stat からのデータのダウンロード方法は，参考としてこの章の付録 12 に載せました．

2　Index 関数を利用して任意の系列を取り出す

ワークシート「dataset」の列について，続く節で，1 つの系列（支出額や品目）を取りだして分析したり，2 つの系列の関係を分析していきます．

そこで，指定した 1 つの系列を抜き出した表を作成し，分析を進めます (ana1)．同様に指定した 2 つの系列を抜き出した表を作成・分析をします (ana2)．また，系列の番号の指定を変更すれば，自動的に対応する系列で分析されるようにします．

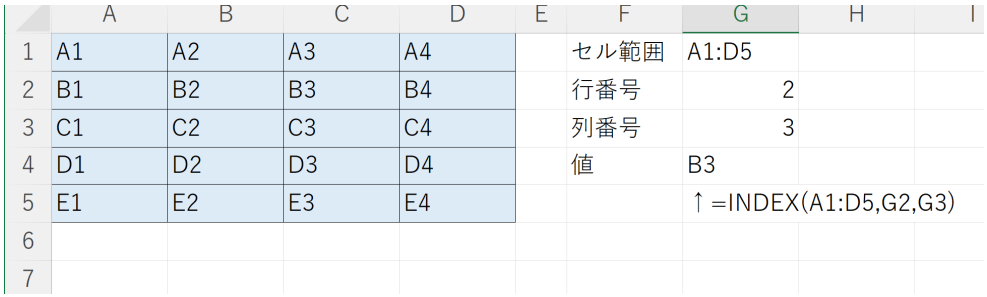
2.1　INDEX 関数

説明のファイル，index_func.xlsx は，教科書のページからダウンロードできます．index 関数は，

INDEX(セル範囲, 行番号, 列番号)

と記述し，その指定されたセルの値を返します．セルの範囲を指定します．図 2 は，index 関数の図解です．

- G4 に，「=INDEX(A1:D5,G2,G3)」の値を表示しています．
- 範囲 A1:D5 は，INDEX で指定する範囲です．
- G2 に行番号を記述し，G3 に列番号を記述し，
- G4 にその行番号と列番号で指定された位置の値を INDEX 関数を使って求めています．



	A	B	C	D	E	F	G	H	I
1	A1	A2	A3	A4		セル範囲	A1:D5		
2	B1	B2	B3	B4		行番号	2		
3	C1	C2	C3	C4		列番号	3		
4	D1	D2	D3	D4		値	B3		
5	E1	E2	E3	E4			↑=INDEX(A1:D5,G2,G3)		
6									
7									

図 2　index 関数

2.2　1 系列分析用のワークシート (ana1) の作成

図 3 のように，セル C1 の値を変更するとその番号の系列の値に自動的に変更されるワークシートを作成します．

(1) 整理されたワークシート (dataset) の B1:G1（図 1）に，消費支出を 1，米を 2,. . ., チョコレートを 6 とする番号を振ってあります．
(2) 1 系列の分析用のワークシート「ana1」を新たに作成します (図 3)．
(3) B 列に地域区分をスピルを使って複写します．

	A	B	C	D
1	行番号	列番号	1	
2	1	地域区分	消費支出	
3	2	全国	3348287	
4	3	札幌市	3220749	
5	4	青森市	2941401	
6	5	盛岡市	3274937	
7	6	仙台市	3410054	
8	7	秋田市	2973311	

	A	B	C	D
1	行番号	列番号	3	
2	1	地域区分	パン	
3	2	全国	31353	
4	3	札幌市	27609	
5	4	青森市	25662	
6	5	盛岡市	28638	
7	6	仙台市	30170	

図 3　1 つの系列を取りだしたワークシート (ワークシート名:ana1)

（a）「ana1」の B2 をクリック

（b）= を入力し，「dataset」の　A2:A55 を範囲指定して，Enter

B2: =dataset!A2:A55

(4) A 列に行番号を設定します．地域区分の行が 1 になるように番号を振っていきます．

（a）ここもスピルを使います．

（b）A2 に次の計算式を設定します．A2: =ROW(B2#)-1

（c）A2 をクリックし，

（d）=ROW(　　と入力

（e）B2 をクリック

（f）#　　を入力．この# で，B2 を含むスピル範囲が指定されます（ここの場合，B2:B55)

（g）) を入力し，Enter

■Spill:Index 関数を使って dataset から値を取り出す

図 3 のように Index 関数を使って，ワークシート ana1 の C1 で指定した列の値をワークシート dataset から抽出していきます．

Index 関数: =index(セル範囲，行番号，列番号)

- ワークシート ana1 の　C2: =INDEX(dataset!B2:G55,'ana1'!A2:A55,'ana1'!C1)
- セル範囲:dataset!B2:G55　　(ワークシート dataset の　B2:G55)
- 行番号:B 列で指定した値なので，スピルを使う場合範囲 ana1　の A2:A55　とします．
- 列番号: 固定した ana1 の　C1 とします．

動画:1 系列を取り出す（Spill）, 音声なし

■絶対参照:Index 関数を使って dataset から値を取り出す

C1 の値を元に，INDEX 関数を使って，dataset から値を取り出します．

- C1: =INDEX(dataset!$B$2:$G$55,A2,$C$1)
- セル範囲は絶対参照にします．列番号は，A2 にします．
- 行番号（A2）は変化するので，相対参照 ($をつけない）にします．
- 複写元: C2 , 複写先: C3:C55

動画:1 系列を取り出す（絶対参照）

2.3　2 系列分析用のワークシート (ana2) の作成

2 系列も使用しますので，図 4 のように新しいワークシート「ana2」に 2 系列のデータを取り出しましょう．ワークシート「ana1」をコピーして「ana2」を作成し，D 列を設定すれば簡単に作成できます．(D2 の計算式は，再度設定したほうが簡単にできます)．

	A	B	C	D	E
1	行番号	列番号	1	2	
2	1	地域区分	消費支出	米	
3	2	全国	3348287	21862	
4	3	札幌市	3220749	27371	
5	4	青森市	2941401	20934	
6	5	盛岡市	3274937	22927	
7	6	仙台市	3410054	21594	

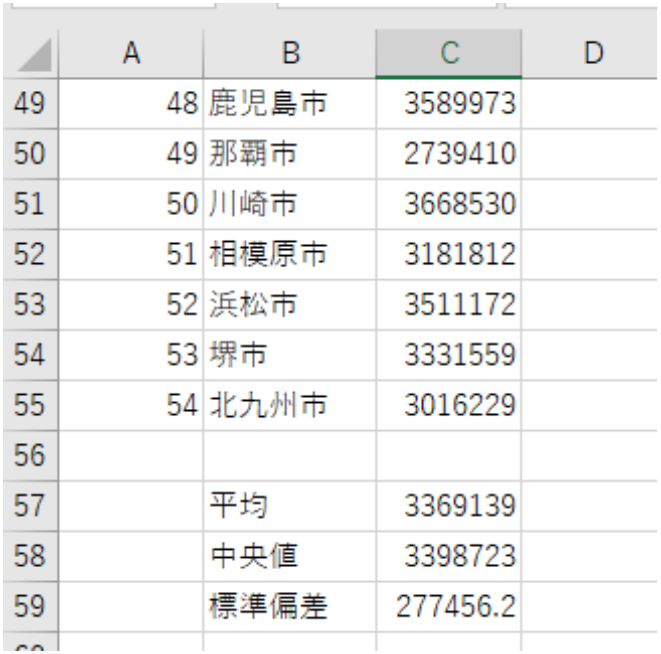
図 4　2 つの系列を取りだしたワークシート (ワークシート名:ana2)

3　平均値，中央値，標準偏差，Z 値 (偏差値)，ヒストグラム

3.1　平均値，中央値，標準偏差

各県庁所在都市，政令指定都市を 1 つの単位と見なし，それらの平均値，中央値，標準偏差，Z 値などを求めます (図 5)．標準偏差は，データの散らばり具合を示す指標で，各サンプル（それぞれの都市）の値と平均値との差（偏差といいます）の大きさを表す値で，偏差の 2 乗の平均値の平方根で求めます．標準偏差は，平均値から平均してどれくらい離れているかの

指標で，大きければ散らばりが大きく，小さければ散らばりは小さいことになります．



	A	B	C	D
49	48	鹿児島市	3589973	
50	49	那覇市	2739410	
51	50	川崎市	3668530	
52	51	相模原市	3181812	
53	52	浜松市	3511172	
54	53	堺市	3331559	
55	54	北九州市	3016229	
56				
57		平均	3369139	
58		中央値	3398723	
59		標準偏差	277456.2	

図 5　平均値，中央値，標準偏差

平均値などを求めるにあたって，全国は含めません．全国は，各都市の値を含んでいるので 2 重に計算することになります．図 5 のように平均値 (関数 AVERAGE)，中央値 (関数 MEDIAN)，標準偏差 (関数 STDEV) を求めましょう．ただし，標準偏差は，「=STDEV(C4:C55)」のように，関数 STDEV は括弧内（引数）に標準偏差を求める範囲を記述します．また，「ana2」の 2 系列のワークシートの各系列の平均値，中央値，標準偏差を求めましょう．

ここで求めた各都市の平均値と全国の値は少し異なります．全国の値は，全世帯の平均（データサイエンス入門で学習し

た加重平均に近い値）であり，各都市の平均値は，各都市を 1 つの単位としてその平均値を求めたもので，各都市の平均値は，全国の値と比べて，世帯数の少ない都市の影響が大きくなります．

3.2　Z 値（偏差値）

Z 値は，平均値と標準偏差を元にどれくらい高い値なのか低い値なのかを示す指標です．偏差値は，Z 値とほぼ同じ目的の値で，100 点満点の試験をイメージできるように変換された値です．

■Z 値，偏差値の考え方　英語と国語で同じ平均点 60 点で，両科目とも 70 点だったとき，どちらの科目がよい得点でしょうか？ この場合，散らばり具合（標準偏差）を考慮します．英語の標準偏差が 10 点で国語が 20 点だとすると，英語は，標準偏差の 1 倍良い点数で，国語は，標準偏差の 0.5 倍良い点数なので英語の方が良い点数になります．

Z 値は，0 で平均点，＋（正）の値で，標準偏差の何倍良い値か，－（負）の値で，標準偏差の何倍悪い値かを示します．

$$\text{Z 値} = \frac{\text{得点 (評価値)} - \text{平均値}}{\text{標準偏差}}$$

で計算します．例の場合，英語の Z 値は 1，国語は 0.5 になります．

このような変換をすることにより，異なる系列（英語と国語の得点）間の値でもある程度比較できるようになります．

偏差値は，Z 値を 10 倍して，50 を加えた値です．

$$\text{偏差値} = \text{Z 値} \times 10 + 50 = \frac{\text{得点 (評価値)} - \text{平均値}}{\text{標準偏差}} \times 10 + 50$$

平均点の Z 値は 0 なので平均点の偏差値は 50 になります．例の場合，英語の偏差値は 60，国語は 55 になります．

■Z 値，偏差値の計算　図 6 のように，ワークシート ana1 の D 列に Z 値，E 列に偏差値を計算しましょう．

(1) 全国の行の Z 値，偏差値は計算しません．

	A	B	C	D	E
1	行番号	列番号	1		
2	1	地域区分	消費支出	Z値	偏差値
3	2	全国	3348287		
4	3	札幌市	3220749	-0.53482	44.65178
5	4	青森市	2941401	-1.54164	34.5836
6	5	盛岡市	3274937	-0.33952	46.60481
7	6	仙台市	3410054	0.147466	51.47466
8	7	秋田市	2973311	-1.42663	35.73369
9	8	山形市	3856930	1.758084	67.58084

図 6　Z 値，偏差値 (ワークシート名:ana1)

(2) Z 値の計算

Spill の利用　C 列の Z 値の部分 D4:D55 をまとめて計算します.

D4:　=(C4:C55 - C57)/C59

絶対参照の利用　C57 の平均値と C59 の標準偏差は複写してもいつも C57，C59 なので，$を付けて絶対参照にします.

D4:　=(C4-$C$57)/$C$59

複写元:　D4　　複写先:　D4:D55

(3) 偏差値の計算

複写の利用

E4: =D4*10+50

複写元: D4:E4　複写先: D4:E55

Spill の利用 (参考)　複写を用いず，スピルを利用してまとめて計算式を設定することもできます．E 列の偏差値の部分 E4:E55 をスピルを使ってまとめて計算します．

E4: = D4:D55 * 10 + 50

E4 の計算式は「=D4#*10+50」と表示されます．D4 の後ろの「#」は，スピル範囲演算子と呼ばれ，D4 を含むスピル範囲（この場合は D4:D55）を指します．

動画:Z 値，偏差値の計算 (Spill)

動画:Z 値，偏差値の計算 (絶対参照)

■2 系列の計算　2 系列の「ana2」の E 列に 1 つめの系列の Z 値，F 列に 1 つめの系列の偏差値，G 列に 2 つめの系列の Z 値，H 列に 2 つめの系列の偏差値を計算してみましょう．

3.3　度数分布表，ヒストグラム (自動で生成)

Excel では，自動でヒストグラムを作成する機能があります．

■ヒストグラム　Excel では，図 7 のように，階級（範囲）をいくつか自動的に設定し，各階級にいくつのサンプルがあるかを表すヒストグラムを作成します．「ana1」の単一の系列の分析で，C1 を 1 にした消費支出の例で説明します．

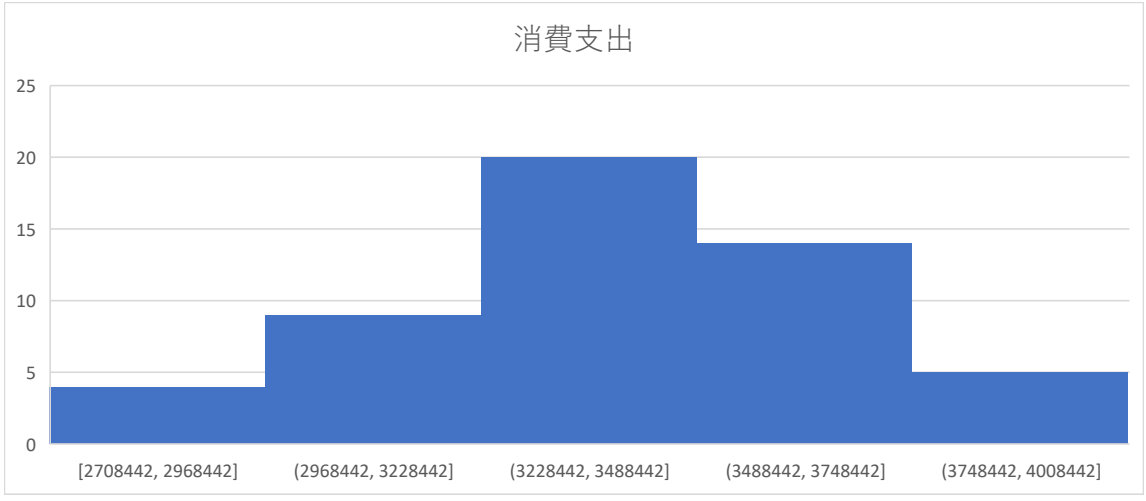


図 7　消費支出のヒストグラム

(1) 度数を数える範囲を指定します．例の場合，札幌市から北九州市までの消費支出を範囲指定します．全国は含めません．
　範囲指定：C4:C55
(2) メニューの 挿入 → グラフ の 統計グラフ → 左上の ヒストグラム
(3) 文字が表示されるよう，グラフを左右に拡大

動画: ヒストグラムの作成

ヒストグラムのタイトルをセルの値で参照

C1 の値を変更することでヒストグラムにする系列を変更できます．ヒストグラムにも系列名 (C2 の値）を表示すると分か

りやすくなります.

(1) グラフタイトルが表示されていなければ，表示させます
　(a) グラフを選択して，グラフ右上の[+]をクリック
　(b) [グラフタイトル]にチェックを入れる
(2) [グラフタイトル]の部分をクリック
(3) 数式バーをクリックして，(=)を入力し，マウスで [C2] をクリック (注意：C2 の選択は，マウスで行います)
　[='ana1'!$C$2] のように表示されます
　うまく行かない場合，グラフのタイトルの編集モードになっていることがあります．その場合は(ESC)キーを押して解除します.
(4) C1 の値を変更すれば，グラフとグラフタイトルが変化します.

セルの値で参照がうまく行かないとき

Mac や一部の Windows マシンで，数式バーをクリックして (=) を入力しても，数式バーに「=」が表示されないことがあるようです．この場合，次のようにすると設定できることが多いようです.

(1) グラフタイトルなどのボックスを削除します.
(2) あらたに，テキストボックスを作成します（「グラフタイトル」ではありません）.
　[メニュー] → [挿入] → [図形] → [テキストボックス]
(3) テキストボックスの中をクリック
(4) 数式バーをクリックして，(=)を入力し，マウスで [C2] をクリック

ヒストグラムの説明

- [2708442, 2968442] は，2708442 以上,2968442 以下を示しています．グラフから 2708442 以上,2968442 以下の度数 (件数) は，4 であることを示しています．
- (2968442, 3228442] は，2968442 より大きく,3228442 以下を示していて，その度数は 9 であることを示しています．2968442 のサンプルは，前の階級 (区間 [2708442, 2968442]) にカウントされます．
- 「(」は「より大」，「)」は「未満」，「[」は「以上」，「]」は「以下」を表しています．
- いくつにわけるか（階級数），階級の範囲は，自動で設定されます．

データがどのように分布しているのかをみるには，このヒストグラムで

- 中心に山があるグラフ（釣り鐘型，例：列番号 1），
- 左右に 2 つの山があるグラフ（2 極化），
- 左に山が高くなり，右に裾が長いグラフ（例：列番号 2）
- 右に山が高くなり，左に裾が長いグラフ（例：列番号 6）

参考：なるほど統計学園，ヒストグラム

■度数分布表　図 7 の各階級の度数を数えて度数分布表を作成することも可能です (図 8)．また，図 7 のグラフを右クリックして データラベルの追加 で，度数を表示することも可能です．

きりの良い階級でのヒストグラムの作成方法は参考としてこの章の付録 13 に載せました．また，Excel の 分析ツール を使って作成することもできます．

階級区間	度数
2708442以上2968442以下	4
2968442より大3228442以下	9
3228442より大3488442以下	20
3488442より大3748442以下	14
3748442より大4008442以下	5

図 8　度数分布表の例

4　2 系列の散布図，相関係数を計算

2 つの系列間で，どのような関係があるのかを分析します．グラフは，散布図 (XY グラフ) を作成します．

■2 系列の散布図　2 系列の散布図は，次のように作成します．

(1) シート「ana2」を選択　(「ana2」は，第 2.3 節で作成済み)
(2) (全国を除く) 各都市の 2 つの系列を範囲指定
　　範囲指定: C4:D55
(3) メニューの 挿入 → グラフ → 散布図 (XY グラフ) またはバブルチャートの挿入 → 散布図 (左上)

C1，D1 を変更すると，系列が変更され，自動的にその系列の散布図が表示されます．
動画:2 系列の散布図

■相関係数の計算　2 つの系列がどのような関係にあるかを見るために，相関係数を求めます．相関係数は，X 軸の値が大きいとき，Y 軸の値も大きいという関係がある場合は 1（または 1 に近い値），X 軸の値が大きいとき，Y 軸の値が小さいという関係がある場合は −1（または −1 に近い値），そのような関係がないとき 0（または 0 に近い値）になるように意図した値です．

相関係数を J3 に求めてみましょう．散布図と同様，全国は含めません．

(1) シート「ana2」を選択
(2) J3 =CORREL(C4:C55,D4:D55)

C4:C55 と D4:D55 は，系列 1 と 2 のデータの範囲です．C1,D1 の値を変更すれば，自動的に系列 1 と 2 のデータが変更され，相関係数も変化します．
動画: 相関係数の計算

■散布図に軸ラベル，相関係数を表示　X 軸と Y 軸が何を表しているのか（軸タイトル），その散布図の相関係数がいくつかなのをグラフに表示します．相関係数の表示は，グラフタイトルのエリアを使用します．J4 にその文字列を生成しておきます．

- J4 =”相関係数:”&J3

散布図の修正:

(1) グラフタイトルが表示されていなければ，表示させます

(a) グラフを選択して，グラフ右上の[+]をクリック

(b) [グラフタイトル]と[軸ラベル]にチェックを入れる

(2) [X 軸]の部分をクリックし，数式バーをクリックして，(=)を入力し，マウスで [C2] をクリック ([='ana2'!$C$2] のように表示されます).

(3) [Y 軸]の部分をクリックし，数式バーをクリックして，(=)を入力し，マウスで [D2] をクリック

(4) [グラフタイトル]の部分をクリックし，数式バーをクリックして，(=)を入力し，マウスで [J4] をクリック

(5) C1,D1 の値を変更すれば，グラフが変化します.

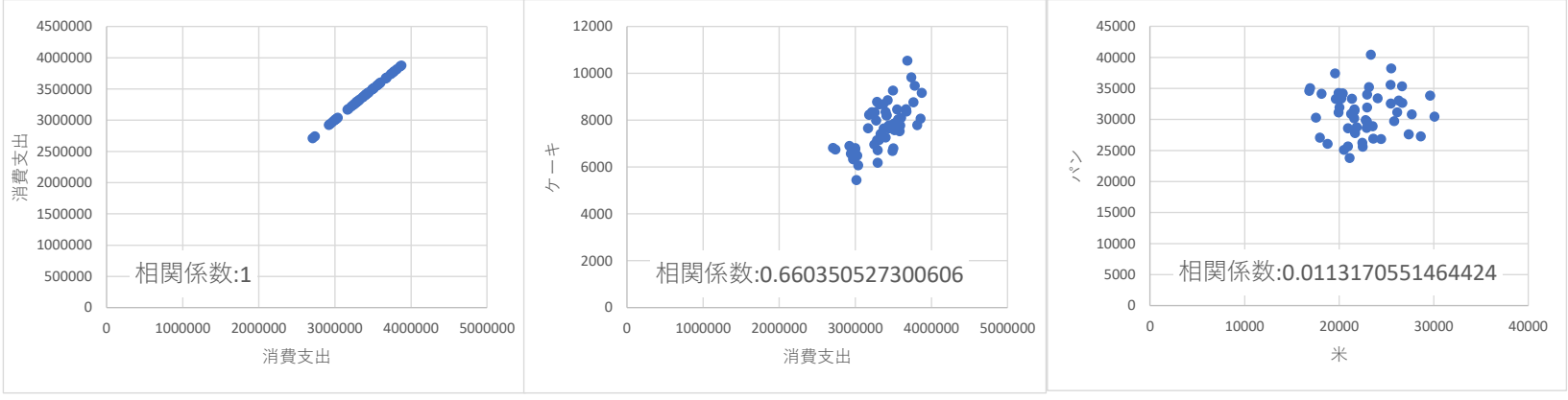


図 9　3 つの散布図と相関係数

図 9 は，C1,D1 の値を変更して，散布図，相関係数を求めたものです.

図 9 左 (C1:1（消費支出）, D1:1（消費支出）)　同じ系列（消費支出）の値で C 列と D 列の値は等しいです．あたりまえですが，C 列の消費支出の値が大きければ，D 列の消費支出の値は大きくなり，散布図の点は，左下から右上への直線上に並びます．このような関係のとき，相関係数は 1 になります．

図 9 中央 (C1:1（消費支出）, D1:4（ケーキの支出）)　消費支出が多い都市は，ケーキの消費支出は大きくなる傾向が読み取れます．左下から右上への直線上に近い関係（正の相関関係）です．この場合，相関係数は 0.66 になります．他にも，(C1:1,D1:5),(C1:1,D1:6),(C1:4,D1:6) などに，正の相関関係があります．

図 9 右 (C1:2（米）, D1:3（パン）)　米の支出が多い都市でも，パンの支出が多いとも少ないともいえないようです．この場合，相関係数は 0.01 と 0 に近い値になります．このように散布図の点が，大きな円に収まっているような関係を無相関といいます．

この例では出ませんでしたが，散布図の点が，左上から右下への直線上に近い位置にあるとき，相関係数は負の値で，その場合負の相関関係があるといいます．正の相関関係では 1 に近いほど，負の相関関係では -1 に近いほど，相関関係は強くなります．

試しに C1，D1 の値を変えて，散布図や相関係数がどうなるか見てみましょう．

5　相関係数行列（相関図行列）

■相関係数行列の作成　例題では，6 個の系列があります．各系列間の相関係数を求め，図 10 のような表の形にまとめたものを相関係数行列と呼びます．左上から右下にかけては，同じ系列同士の相関係数ですので 1 になります．また，(系列 A と系列 B の相関係数) と (系列 B と系列 A の相関係数) は同じであるので，左上から右下に 1 が並ぶ対角線を軸として，対称な位置の値は等しくなります．

作成方法は，ここでは愚直に，一つずつ，2 つの系列の相関係数を求めていきます（統計ソフトウエア（R など）では自動で計算できるものが多いようです．また，Excel でも計算式の複写で作成する方法もありますが，計算式が複雑になるため

	消費支出	米	パン	ケーキ	せんべい	チョコレート
消費支出	1.00	-0.04	0.30	0.66	0.45	0.61
米	-0.04	1.00	0.01	-0.06	-0.06	-0.02
パン	0.30	0.01	1.00	0.36	0.08	0.39
ケーキ	0.66	-0.06	0.36	1.00	0.21	0.54
せんべい	0.45	-0.06	0.08	0.21	1.00	0.18
チョコレート	0.61	-0.02	0.39	0.54	0.18	1.00

図 10　相関係数行列

省略します).

(1) 図 10 の表頭と表側の系列名を入力しておきます.
(2) 同じ系列名が示すセル（左上から右下への対角線上のセル）に 1 を入力します.
(3) 消費支出 (1) と 米 (2) の相関係数を求め，記入します.
　（a）相関係数を計算するワークシート「ana2」に移動します.
　（b）列番号 (C1) に消費支出の 1 を入力します.
　（c）列番号 (D1) に米の 2 を入力します.
　（d）自動で 2 つの系列の相関係数が計算されるので，その計算結果 F3 の値をコピーします.
　（e）相関係数行列の対応する部分（消費支出と米が交わる部分）2 カ所に「値貼り付け」で貼り付けます.
(4) 他の組み合わせについても同様に作業をしていきます（15 個の相関係数を計算).

動画: 相関係数行列の作成

■相関係数行列のヒートマップ　図 10 では数字が羅列された表であり，直感的に理解するため，図 11 のようなヒートマップグラフを作成します．ヒートマップは，値の大きさにより，背景色を変化させます．相関係数は，正の相関 (+1), 負の相関 (−1), 無相関 (0) の 3 つの極があるので，3 色の強さで色分けします．ここでは，正の相関を赤, 負の相関を緑，無相関を黄色にし，それらの中間の値の場合，中間色（グラデーションカラー）になるように設定しました．

	消費支出	米	パン	ケーキ	せんべい	チョコレート
消費支出	1.00	-0.04	0.30	0.66	0.45	0.61
米	-0.04	1.00	0.01	-0.06	-0.06	-0.02
パン	0.30	0.01	1.00	0.36	0.08	0.39
ケーキ	0.66	-0.06	0.36	1.00	0.21	0.54
せんべい	0.45	-0.06	0.08	0.21	1.00	0.18
チョコレート	0.61	-0.02	0.39	0.54	0.18	1.00

図 11　相関係数行列のヒートマップ

ヒートマップグラフの作り方

(1) 色付けたい部分を範囲指定します（表頭や表側は含めません）.
(2) メニューの [ホーム] → [条件付き書式] → [新しいルール]
(3)「新しいルール」のウインドウが表示されます（図 12）.
　ルールの種類:　セルの値に基づいてすべてのセルを書式設定
　書式スタイル:　3 色スケール　（3 つの色を入力するようになります）
　最小値の列:　負の相関の色を設定
　　[種類:] 数値　[値:] −1 [色:] 緑色

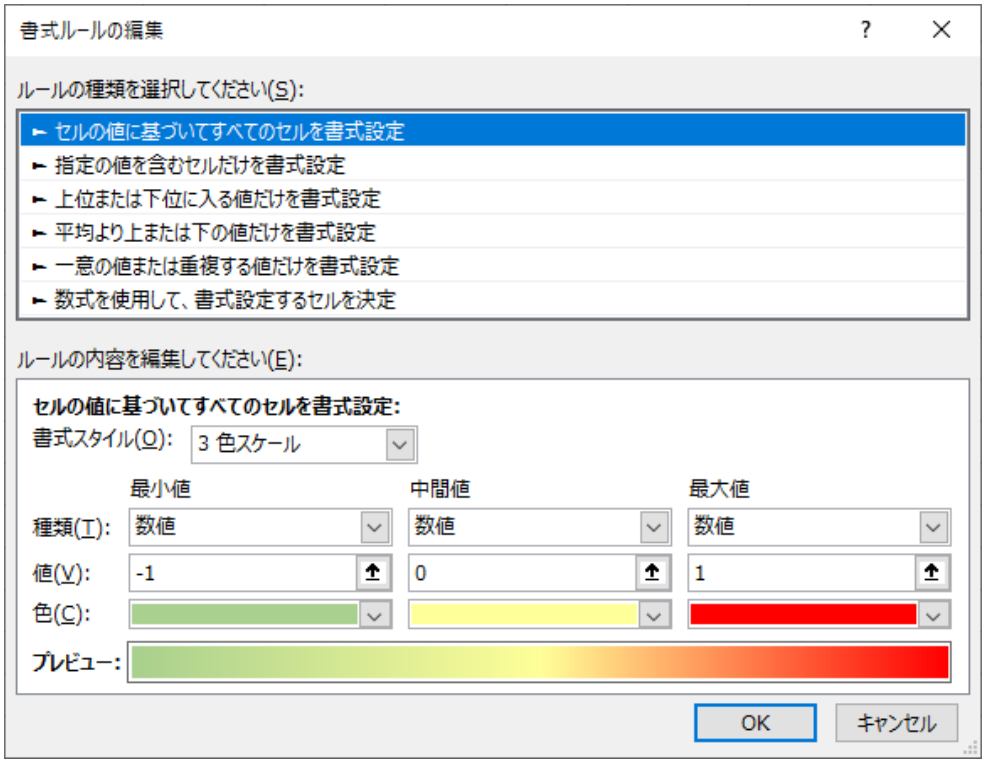


図 12　相関係数行列のヒートマップの設定

中間値の列:　無相関の色を設定

種類: 数値　値: 0　色: 黄色

最大値の列:　正の相関の色を設定

種類: 数値　値: 1　色: 赤

動画: 相関係数行列のヒートマップの作成

■相関図行列　図 13 のように，散布図，ヒストグラムと相関係数の図を表のように表示したものは相関図行列と呼ばれています（これは例で，さまざまな種類の相関図行列があります）．この図は，各系列間の全体の傾向をみるのに使われます．したがって，詳細の部分を見るときは，各 2 系列の散布図，各系列のヒストグラムを参照します．ヒストグラムで各系列の大まかな分布を読み取り，各散布図で大まかな相関関係，両図から外れ値（他の値から大きく離れているサンプル）の存在を見ます．

相関図行列は，Excel での作成は大変で，統計ソフトウエア（R など）を使用した方が良いでしょう．

■Excel での相関図行列の作成（参考）

(1) 新しいワークシートで，列幅，行の高さを調整し，大きな正方形のセルにします．
(2) 1 行目，1 列目，最後の行，最後の列の列幅，行の高さを調整し，系列名を入力します．
(3) シート「ana1」で，列番号を変更して，ヒストグラムを作成し，その図を正方形にします．
(4) ヒストグラムをコピーします．
(5) 相関図行列のワークシートで，ヒストグラムを　対応するセルに図（拡張メタファイル）で貼り付けます．
(6) セルの罫線に重ならないように大きさを調整します．
(7) 各ヒストグラムを貼り付けていきます．
(8) シート「ana2」で，列番号 2 つを変更して，散布図を作成し，その図を正方形にします．
(9) 相関図行列のワークシートで，散布図を　対応するセルに図（拡張メタファイル）で貼り付けます．
(10) セルの罫線に重ならないように大きさを調整します．
(11) 各散布図を貼り付けていきます．
(12) 右上の相関係数を貼り付けていきます．

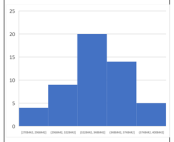
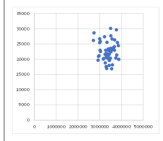
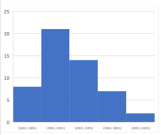
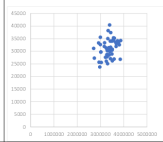
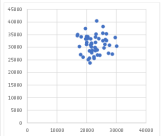
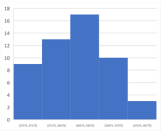
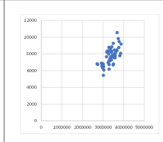
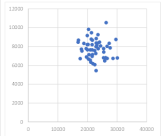
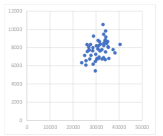
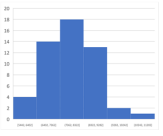
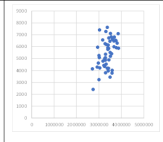
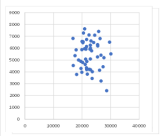
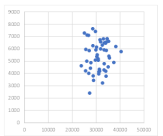
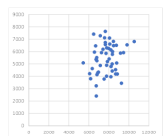
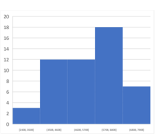
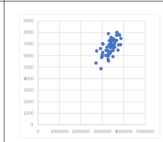
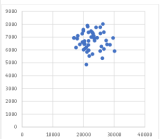
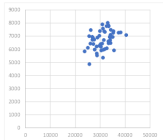
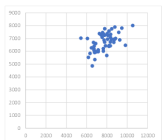
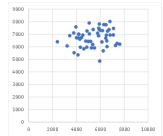
	消費支出	米	パン	ケーキ	せんべい	チョコレート	
消費支出		-0.04	0.30	0.66	0.45	0.61	消費支出
米			0.01	-0.06	-0.06	-0.02	米
パン				0.36	0.08	0.39	パン
ケーキ					0.21	0.54	ケーキ
せんべい						0.18	せんべい
チョコレート							チョコレート
	消費支出	米	パン	ケーキ	せんべい	チョコレート	

図 13　相関図行列

動画: 相関係数行列の作成

6　相関関係と因果関係

相関関係は，一方の値が大きければもう片方の値も大きいという関係 (正の相関関係) という直線的な関係（散布図では，右上がりの直線に近い関係），または，一方の値が大きければもう片方の値が小さいという関係 (負の相関関係) という直線的な関係（散布図では，右下がりの直線に近い関係），を表します．しかし，相関関係がある場合でも，どちらかが原因で，もう片方が結果という因果関係があるとは限りません．

この節の分析では，家計は支出できる額の合計（消費支出) を決め（原因），その後，どの品目にどれくらい支出するかを決定する（結果）とします．各品目への支出は家計は支出できる額の合計を考えずに，各品目へ適当に支出し（原因），その支出額の合計が決まる（結果）とは考えないとします．

図 11 をみると，チョコレートとケーキの相関係数は 0.54 と比較的大きな値で正の相関関係があります．これは，図 13 のチョコレートとケーキの散布図を見ても右上がりの直線に近い傾向が読み取れます．

ここで，チョコレートの消費支出が多いことが原因で，その結果としてケーキの消費支出が多くなるという因果関係が考えられるでしょうか？ また，逆にケーキの消費支出が多いことが原因で，その結果としてチョコレートの消費支出が多くなるという因果関係が考えられるでしょうか？ 洋菓子好きは，ケーキ好きでチョコレート好きのことが多いという相関関係はありますが，どちらかが原因となるとは考えにくいと思います．

チョコレートとケーキの消費支出の共通の原因を考えられないでしょうか？ 図 11 や図 13 と消費支出とケーキおよびチョコレートとの相関係数は，それぞれ 0.66 と 0.61 であり，正の相関関係があります．チョコレートとケーキは嗜好品と考えられるので，消費支出水準（または所得水準）が高いとチョコレートとケーキへの消費支出が増えると考えられないでしょうか？

そうだとすると，消費支出が原因でケーキの消費支出が結果という因果関係，消費支出が原因でチョコレートの消費支出

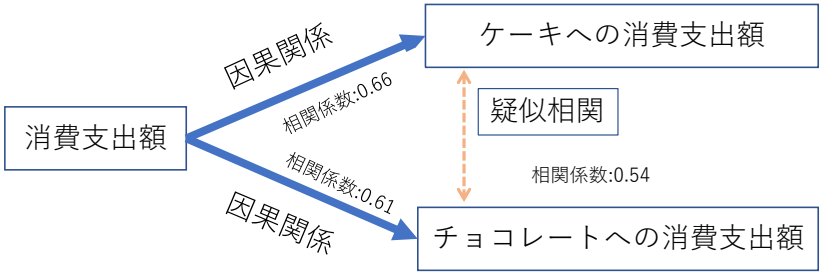


図 14　疑似相関

が結果という因果関係が考えられます（図 14）．この結果として，ケーキとチョコレートのあいだに相関関係が生じました．このように，別の共通の原因によって相関が生じることを疑似相関と呼びます．

7　ソート・順位

各系列の値を大きい順（降順）や小さい順（昇順）に並べ替え，どのようなサンプル（都市）が上位に来るのかをみることにより，その系列の特徴を分析することが出来ます．また，2 つ以上の系列の順位を比較することにより，系列間の関係を考察することができます．そこで，図 15 のような表を作成します．

7.1　準備

(1) Excel に新しいワークシートを作成します．ワークシート名は，例えば ソート・順位 とします．
(2) 図 15 の 1 行目から 3 行目を入力します．ワークーシート dataset からコピーアンドペーストを使うと楽です．
(3)「金額」と「順位」は， セルを結合して中央揃え を使っています．

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ソートの基準列		2										
2		金額						順位					
3	地域区分	消費支出	米	パン	ケーキ	せんべい	チョコレート	消費支出	米	パン	ケーキ	せんべい	チョコレート
4	東京都区部	3,873,418	19,942	34,260	9,167	5,874	7,481	1	44	9	5	23	7
5	山形市	3,856,930	24,447	26,818	8,064	7,094	6,946	2	14	46	21	5	21
6	富山市	3,813,641	25,462	32,557	7,784	6,297	7,768	3	12	21	25	15	4
7	さいたま市	3,780,012	21,368	33,327	9,469	5,915	7,814	4	33	17	3	21	3
8	新潟市	3,764,133	29,639	33,839	8,757	6,509	6,930	5	2	14	8	12	24
9	千葉市	3,737,810	20,389	34,177	9,824	6,542	6,474	6	39	10	2	11	32
10	金沢市	3,684,997	26,316	33,036	10,539	6,809	8,016	7	8	19	1	7	1

図 15　ソート・順位用の表

7.2　ソート

ソートは，ワークシート dataset の 4 行目から 55 行目までとします．「全国」は，並べ替えや順位の対象とはしないので，ソートの範囲には含めません．ソートは，sort 関数を使って，ワークシート dataset の範囲を基にして，並べ替えをして，ワークシート ソート・順位 に表示させます．Sort 関数は，データサイエンス入門の第 5 章の 5-18 付近に説明があります．

(1) ソートの基準の列を C1 に入力していきます．図の例では，「消費支出」を基準にするので，2 と入力しておきます．

(2) A4 に sort 関数の計算式を入力します．

（a）=SORT(と入力します．

（b）ソートの範囲を指定するので，dataset をクリックして，A4 から G55 までを範囲指定します．

（c）, を入力し，基準の列が書かれた，ソート・順位 の C1 をクリックします．

(d) 大きい順（降順）にするので，[,]を入力し，[降順]を選択します．

(e) 行を並べ替えるので，[,]を入力し，[行で並べ替え]を選択します．

(f) [)]を入力し，(Enter)を入力します．

(3) ソート結果が表示されたと思います．

(4) A4 は，[=SORT(dataset!A4:G55, ソート・順位!C1,-1,FALSE)]となっていると思います．

(5) C1 を，$2, 3, \ldots, 7$ と変化させると異なる基準で並べ替えが行われます．

動画: ソート

7.3 順位

図 15 右のように，各都市の順位を求めてみましょう．順位もデータサイエンス入門で学習した RANK 関数（スピルを利用）を使います．データサイエンス入門の第 5 章の「7 RANK 関数を利用して順位を求める」に説明があります．

(1)「消費支出」の列の順位を求めます．

(2) H4 をクリックし，[=RANK(]と入力

(3) B4:B55 を指定 (順位を並べ替える値を指定，スピルを利用して下まで求めるために，B4 から B55 まで範囲指定),

(4) [,]を入力し，B4:B55 を指定（順位を求める範囲),

(5) [,]を入力し，0 (降順）を指定

(6) [)]を入力し，(Enter)を入力します．

(7) H4 は，[=RANK(B4:B55,B4:B55,0)]となっていると思います．

(8)「米」,「パン」,... の列も同様に順位を設定するので,「消費支出」の列の計算式を複写します．

(9) H4:H55 の範囲を右に複写します．
　　複写元: H4:H55 , 複写先: I4:M55

動画: 順位

系列間の関係が見えてくると思います．基準列を 5(ケーキ) としてみると，ケーキの上位には，チョコレートの順位も高い（小さい）都市が出てきます．しかし，完全に一致しているわけではありません．たとえば，1 位は，ケーキ，チョコレートともに金沢市ですが，2 位は，ケーキは千葉市，チョコレートは鳥取市です．

8　回帰分析

図 9 をみると，消費支出が大きいと，ケーキへの支出が多くなる傾向が見えます．実際，相関係数も 0.66 と大きな値を示しています．また，前節 (7 節) でも，消費支出でソートをすると，上位の都市のケーキの順位は高順位です．また，図 9 では，点の分布が右上がりの直線に近い位置にあります．

このような関係があるとき，消費支出額から，ケーキの支出額を推定することができます．このように直線の関係があると考えて，推定式を求めることを線型回帰分析または単に回帰分析といいます．また，ケーキの支出額は 1 つの変数（系列）である消費支出から推定されるので，単回帰分析（または線型単回帰分析）とよばれます．2 つ以上の変数（系列）から推定する分析は重回帰分析とよばれます．

ここでは，消費支出額からケーキの支出額を推定します．消費支出額は説明変数とよばれ，x という変数で表します．ケーキの支出額は被説明変数とよばれ，y という変数で表します．

$$y = ax + b$$

という式（1 次式，直線）で表し，x に消費支出額を代入するとこの式により，ケーキの支出額の推定値 y が求まります．a

は回帰係数，b は定数項と呼ばれ，各都市の x と y の値から求めることができます．

8.1　準備

回帰分析は，ワークシート ana2 をコピーして利用します．

(1) ワークシート ana2 を右クリックして，移動またはコピー
(2) コピーを作成する にチェックを入れ，OK をクリック
(3) コピーで作成したワークシート名を 回帰分析 にしましょう

E〜J 列は，不要なので削除します．また，散布図は書き直しますので，一度削除します．

(1) E〜J 列を範囲指定
(2) 右クリックして，削除

図 16 のように，H,I,J 列に分析用の表を作成します．

(1) 表頭の転記
H2: =B2　I2: =C2　J2: =D2
(2) K2 に「推定値」と入力
(3) 各系列の値を転記します．ただし，全国の値は分析に含めませんので，4 行目から 55 行目までです．
H3: =B4:B55　I3: =C4:C55　J3: =D4:D55

動画: 回帰分析準備

	A	B	C	D	E	F	G	H	I	J	K
1	行番号	列番号	1	4							
2	1	都市名	消費支出	ケーキ		相関係数		都市名	消費支出	ケーキ	推定値
3	2	全国	3348287	7716		0.660351		札幌市	3220749	8339	
4	3	札幌市	3220749	8339				青森市	2941401	6563	
5	4	青森市	2941401	6563				盛岡市	3274937	7984	
6	5	盛岡市	3274937	7984				仙台市	3410054	8200	
7	6	仙台市	3410054	8200				秋田市	2973311	6337	
8	7	秋田市	2973311	6337				山形市	3856930	8064	
9	8	山形市	3856930	8064				福島市	3512916	7573	
10	9	福島市	3512916	7573				水戸市	3309829	7141	
11	10	水戸市	3309829	7141				宇都宮市	3373212	7649	

図 16　回帰分析データ表

8.2　回帰分析の実行

8.2.1　分析ツールの導入

回帰分析は，Excel の「分析ツール」を使います．分析ツールは，標準では導入されていません．次のように設定します．

(1) Excel のメニューの ファイル → オプション
(2) アドイン を選びます．
(3) 下方の 管理 (A) の右の Excel アドオン のなかの 設定 をクリック
(4) 分析ツール にチェックを入れます（他にも，ソルバーアドイン や 分析ツール - VBA にもチェックを入れておくと便利です）．

(5) OK をクリック

動画: 回帰分析, 分析ツールの導入準備

8.2.2　回帰分析ツールの実行



図 17　回帰分析のオプションの設定

(1) Excel のメニューの データ → 分析の中の データ分析
(2) 分析ツールの中から 回帰分析 を選択し，OK
(3) 回帰分析のオプション設定が表示されるので，図 17 のように設定します．

入力 Y 範囲: 被説明変数の範囲を指定します．表頭も含めます．この例の場合，ケーキの範囲 J2:J54 としラベルにチェックを入れます．

入力 X 範囲: 説明変数の範囲を指定します．表頭も含めます．この例の場合，消費支出の範囲 I2:I54 としラベルにチェックを入れます．

ラベル: 説明，被説明の範囲で表頭を含めていますので，チェックを入れます．

出力オプション 一覧の出力先 を同じシートの場所, 例えば，N2 を設定します．

(4) 図 18 のように N2:V19 に結果が表示されます．

動画: 回帰分析, 回帰分析ツールの実行

8.3 回帰分析結果の分析

8.3.1 推定値の計算

分析結果（図 18）の O18 に推定式の b（定数項）の値，O19 に推定式の a（回帰係数）の値が表示されます．したがって，この分析の場合の回帰式は，$y = 0.002424x - 411.521$ となります．この式を使って，K 列の推定値を計算しましょう．

K3: =I3:I54*O19+O18

大きく外れている都市もありますが，だいたい近い値が示されていると思います．この推定値の点 ((消費支出額, 推定値)) を結んだ線が，近似曲線になります．

8.3.2 グラフの作成

回帰式の直線 (近似曲線）を含めた図 18 のようなグラフを作成してみましょう．

	地域区分	消費支出	ケーキ	推定値
3	札幌市	3220749	8339	
4	青森市	2941401	6563	
5	盛岡市	3274937	7984	
6	仙台市	3410054	8200	
7	秋田市	2973311	6337	
8	山形市	3856930	8064	
9	福島市	3512916	7573	
10	水戸市	3309829	7141	
11	宇都宮市	3373212	7649	
12	前橋市	3439750	7645	
13	さいたま市	3780012	9469	
14	千葉市	3737810	9824	
15	東京都区部	3873418	9167	
16	横浜市	3601822	8098	
17	新潟市	3764133	8757	
18	富山市	3813641	7784	
19	金沢市	3684997	10539	

概要　　回帰分析によ

回帰統計	
重相関 R	0.660351
重決定 R2	0.436063
補正 R2	0.424784
標準誤差	772.4947
観測数	52

分散分析表

	自由度	変動	分散	観測された分散	有意 F
回帰	1	23071689	23071689	38.66236	1E-07
残差	50	29837406	596748.1		
合計	51	52909095			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-411.521	1317.876	-0.31226	0.756141	-3058.55	2235.51	-3058.55	2235.51
消費支出	0.002424	0.00039	6.217906	1E-07	0.001641	0.003207	0.001641	0.003207

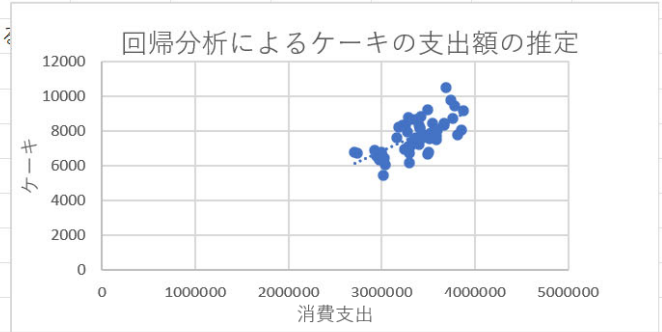


図 18　回帰分析の分析表

(1) グラフ化する範囲（消費支出とケーキ），I2:J54 を範囲指定

(2) メニューの 挿入 → 散布図 → 散布図

(3) グラフ右の ＋(要素の追加) から 近似曲線 → 線形

(4) 点線で回帰直線が表示されたと思います．

(5) グラフのタイトル，軸のタイトルを付けてみます

グラフのタイトル　「回帰分析によるケーキの支出額の推定」のようにしてみます.

（a） P2 = ”回帰分析による”&J2&”の支出額の推定”

（b） グラフのタイトルをクリック

（c） 数式バーをクリックし，(=)を入力し，P2 をクリック

X 軸ラベル　消費支出 (I2) とします.

（a） グラフ右上の + ボタンで，軸ラベル を追加します.

（b） X 軸ラベルをクリックし，数式バーをクリックし，(=)を入力し，I2 をクリック

Y 軸ラベル　ケーキ (J2) とします.

（a） X 軸ラベルをクリックし，数式バーをクリックし，(=)を入力し，J2 をクリック

点線を読み取ることで推定値が分かります．都市によっては，推定値（K 列）がケーキの支出額（J 列）が異なることが，グラフでは，線と点が離れていることに現れています.

動画: 回帰分析, グラフの作成

8.3.3　回帰分析のあてはまりの良さ

回帰分析のあてはまりの良さは，決定係数（Excel の表示では，重決定 R2) を見ます．この値は，0 以上 1 以下の値をとり，1 に近ければ近いほど，あてはまりがよいことを示しています．いくつ以上で，「あてはまりよい」と言えるなどの基準はないのですが，おおよそ 0.8 以上で「あてはまりが非常によい」，おおよそ 0.5 以上で「あてはまりがよい」，0.2~0.3 以上で，「ややよい」と言われています.

9　練習問題

本例題は，消費支出と米，パン，ケーキ，せんべい，チョコレートの分析を行ってきました．ワークシート「全品目」から他の品目のデータを利用可能です．本例題でのワークシートを適切に利用して，他の品目を選択して，以上の節で説明した分析を行い，考察してみましょう．

また，消費支出額を説明変数とし，他の品目の支出額を被説明変数とする線形回帰分析を行ってみて，考察してみよう．

10　階層的クラスター分析（分析過程の学修）

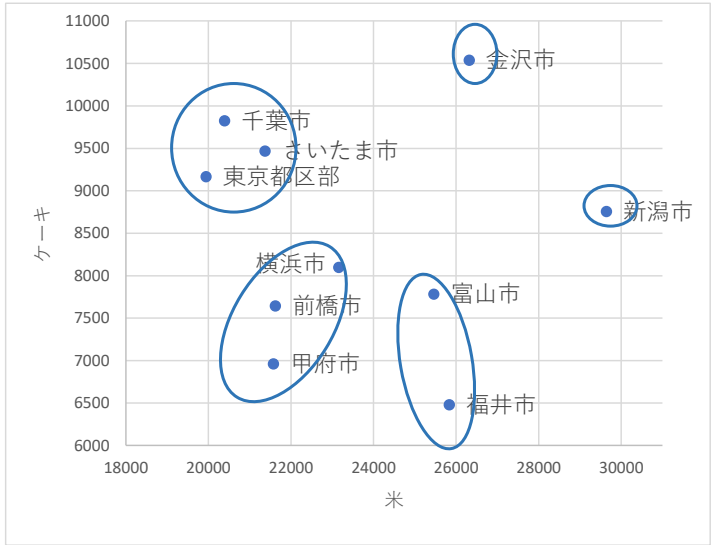
クラスタリング (クラスター分析) とは，サンプル間の類似度に基づいて，サンプルをクラスター（グループ）に分類する方法です．サンプル間の類似度には，サンプル間の距離を用います．距離が近いほど（小さいほど），類似しているとします．ここでの家計調査の分析では，散布図上の 2 都市間の距離を使います．散布図で，近くにある都市を同じクラスターにする分析です．

クラスタリングにはさまざまな方法がありますが，分析結果は，図 19 右のような散布図にグループ分けを記した図や図 19 下のデンドログラム（樹形図）で表現します．

ここでは，家計調査のデータで，10 都市の 2 つの品目 (この例では，米とケーキ）の支出額をもとに，都市をグループ化していきます．図 19 は，データ例（左上）とその散布図表示によるクラスタリング結果（右上），樹形図表現（下）です．

この節では，クラスタリングの全過程を Excel で自動計算するのではなく，学修者が途中での図や表を見ながらどのクラスターを結合するのかを決めていきます．ここでは，分析・計算過程を理解することを学修目標にします．

	都市名	米	ケーキ
A	前橋市	21619	7645
B	さいたま市	21368	9469
C	千葉市	20389	9824
D	東京都区部	19942	9167
E	横浜市	23158	8098
F	新潟市	29639	8757
G	富山市	25462	7784
H	金沢市	26316	10539
I	福井市	25835	6480
J	甲府市	21576	6960



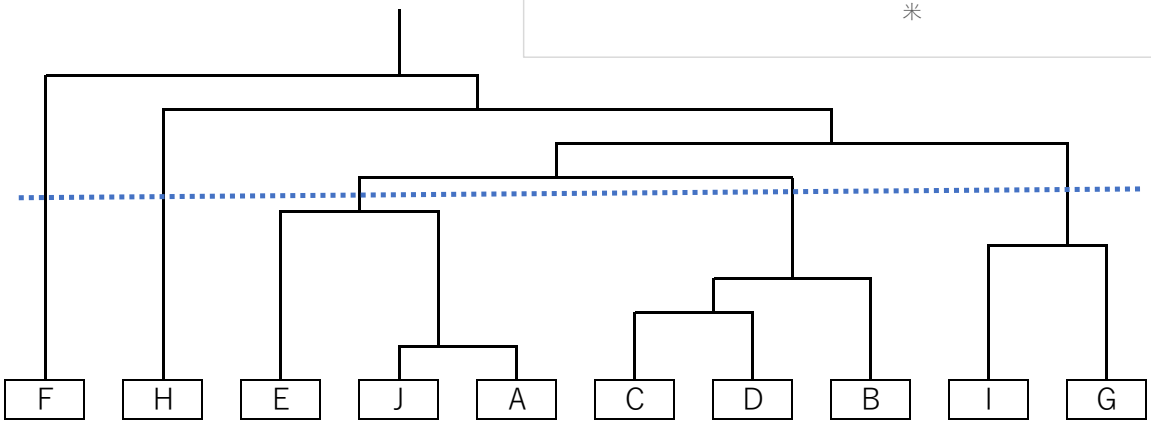


図 19　例題のデンドログラム（樹形図）

10.1　クラスタリングのやりかた

10.1.1　サンプル間の距離

サンプル間の距離は，ユークリッドの距離を用います．2 系列 ((x, y)，(米，ケーキ)) における 2 つの都市 A，B の距離は，次のようになります．

$$\text{A と B の距離} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

となります．たとえば，前橋市 (A) とさいたま市 (B) の (米, ケーキ) の距離は，

$$\text{前橋市とさいたま市の (米, ケーキ) の距離} = \sqrt{(21619 - 21368)^2 + (7645 - 9469)^2} = 1841.189$$

となります．

10.1.2　クラスターの結合方法

初期のクラスターの場合：

- 初期のクラスターは，各サンプルが 1 つのクラスターになります．
- 全クラスター間の距離を求めます．
- 距離が最小のクラスターを結合し，新しいクラスターにします．

図 19 の場合，サンプル J と A の距離が最小の距離なので，最初に結合しています．

複数のサンプルからなるクラスターの結合：

- $\{J, A\}$ からなるクラスター，$\{C, D\}$ からなるクラスターの距離は，各クラスターから 1 つずつサンプルを取り出し，そのサンプル間の距離を求め，そのすべての組み合わせの最小値とします．この場合，(J,C) の距離，(J,D) の距

離，(A,C) の距離，(A,D) の距離の最小値を $\{J, A\}$ と $\{C, D\}$ の距離とします（この方法は，最短距離法と呼ばれています）.

- この作業をくり返していきます.
- 最短距離法は，単純であり，表計算ソフトウエアで簡単に実現できるので，ここでは利用することにしました.

10.2 Excel によるクラスタリングの実習

教科書のページ Clustering_ex1.xlsx に, この例題のファイルが入っています．また，Clustering_calc.xlsx はデータ入力前のファイルで，他のデータで分析するときに利用します．このファイルでは，2 系列，10 サンプルの分析ができます.

次の部分から成り立っています.

入力エリア 1 行目から 11 行目までのエリアで，黄色の部分にデータを入力します（例題のデータはすでに入力されています）.

距離計算エリア 2 つのサンプル間の距離を計算しています．同じクラスターに所属するとき，距離は計算しない（表示は 9999999）ように設定しています．距離の最小値は，赤のセルで示しています.

距離計算エリアのクラスタ (B20:B29) ここに，どのクラスターに所属するのかを記述していきます（手作業）．B20:B29 を書き換えることにより，クラスタを結合していきます.

樹形図作成過程エリア 作成過程を履歴として残していきます.

散布図 各サンプルの点のデータラベルはクラスタを表します．B20:B29 を書き換えると，自動で書き直されます.

作業手順：

(1) 初期の状況で，赤のセルは，A と J の交差する場所です．これは，A と J の距離が最小であることを示しています.

そこで，J を A に結合します．メモとして，B34 に J, B35 に A と記入します．結合の作業は，B20:B29 のうち，J のセルを A に書き換えます．メモとして，このときのクラスタの情報 B20:B29 を 2 回目のクラスタの部分 B36:B45 に値貼り付けをしておきます．

(2) 2 回目では，C と D の交差する場所が，最小の距離です．そこで，D を C に結合します．メモとして，C34 に D, C35 に C と記入します．結合の作業は，B20:B29 のうち，D のセルを C に書き換えます．メモとして，このときのクラスタの情報 B20:B29 を 2 回目のクラスタの部分 C36:C45 に値貼り付けをしておきます．

(3) 3 回目は C を B に結合します．B20:B29 のうち，C のセルを B に書き換えますが，複数あるので注意してください．

(4) 同様にくり返していきます．

(5) 最後（9 回目）で 1 つのクラスタに結合されます．

動画: クラスタリング，結合作業 (音声付き)
動画: クラスタリング，結合作業 (音声なし)

10.3　分析結果の解釈

10.3.1　デンドログラム・散布図

図 19 下のデンドログラムは，どのようにクラスターが結合されていくのかを樹形図（ツリー) で表したものです．一番下の結合した部分が最初の結合で，その上が 2 回目の結合を表しています（本来は，距離が縦軸になるのですが，図 19 では，回数にしています)．デンドログラムは，Excel では簡単には作成できません．

図 19 下のデンドログラムでは，青のラインで，3 つの要素数 2 以上のクラスターと 2 つのはずれた 1 つのクラスターになります (全部で 5 個のクラスター)．このラインで，図 19 右の散布図は，クラスタを青線で囲っています．このラインは，分析の目的などを考えて，設定します．

例題のように，2 次元の散布図 (米とケーキ）もしくは 3 次元の散布図で表現できる場合，距離やサンプルの散布の状況，

クラスタ間の位置関係を直感的に把握できます．本分析では，散布図をみながら分析を進めます．

10.3.2　クラスター数の決定

本分析では，分析の途中で貼り付けた，クラスタの状況と各クラスター数の要素数推移の表（図 20) および 2 次元の散布図を利用します．また，B20:B29 に，各回数のクラスタの推移の部分 (B36:J45) をコピーアンドペーストすることにより，散布図にクラスターの状況が表示され，クラスター数を決める参考情報にします．

10.4　練習問題

「S110.xslx」 のワークシート「全品目」のデータを使い，クラスター分析をしてみましょう．ただし，2 品目，10 都市を選択して，分析します．「Clustering_calc.xlsx (クラスタリング用ファイル（データ入力前）)」の背景が黄色のセルに入力していきます．

	A	B	C	D	E	F	G	H	I	J
32	樹形図作成過程									
33	回数	1	2	3	4	5	6	7	8	9
34	統合元	J	D	C	I	E	B	G	H	F
35	統合先	A	C	B	G	A	A	A	A	A
36		A	A	A	A	A	A	A	A	A
37		B	B	B	B	B	A	A	A	A
38		C	C	B	B	B	A	A	A	A
39		D	C	B	B	B	A	A	A	A
40	クラスタ	E	E	E	E	A	A	A	A	A
41		F	F	F	F	F	F	F	F	A
42		G	G	G	G	G	G	A	A	A
43		H	H	H	H	H	H	H	A	A
44		I	I	I	G	G	G	A	A	A
45		A	A	A	A	A	A	A	A	A
46	クラスター数	9	8	7	6	5	4	3	2	1
47	クラスターの要素数									
48	A	2	2	2	2	3	6	8	9	10
49	B	1	1	3	3	3	0	0	0	0
50	C	1	2	0	0	0	0	0	0	0
51	D	1	0	0	0	0	0	0	0	0
52	E	1	1	1	1	0	0	0	0	0
53	F	1	1	1	1	1	1	1	1	0
54	G	1	1	1	2	2	2	0	0	0
55	H	1	1	1	1	1	1	1	0	0
56	I	1	1	1	0	0	0	0	0	0
57	J	0	0	0	0	0	0	0	0	0

図 20　クラスターの変化

11　COVID-19 の分析 (時系列データの分析)

11.1　データの説明

2019 年から全世界的に流行した感染症である COVID-19 の日本における感染者数, 死者数を分析します．データは,「厚生労働省: データからわかる－新型コロナウイルス感染症情報－」https://covid19.mhlw.go.jp/ から都道府県別，日別の新規陽性

者数と死者数のデータと，人口 100 万人あたりの分析をするため，日本の統計 (https://www.stat.go.jp/data/nihon/index1.html) より日本の都道府県別の人口のデータを利用します（2023/04/18 ダウンロード）．新型コロナウイルス感染症情報には，都道府県別の新規陽性者数と死者数のデータには，人口 10 万人あたりや週別のデータもありますが，これらは学修の過程で計算するため利用しないことにしました．

データは，教科書のページから S20_covid19.xlsx というファイル名でダウンロードできます．このファイルには，3 つのワークシート「都道府県別新規陽性者数」,「都道府県別死者数」,「人口」が含まれています．

2023/4/18 現在で，新規陽性者数は 2020/1/16〜2023/4/18，死者数は 2020/5/9〜2023/4/18 のデータがオープンデータとして公表されています．2 つのデータで，データの開始日が異なることに注意してください．共通のデータの開始日 2021/5/9 から 2022 年度の年度末 2023/3/31 までを全期間とするデータで分析します．

11.2　原系列のグラフ化

このデータは，日別のデータで，時とともに変化するデータなので，時系列データと呼ばれています．時系列データは折れ線グラフなどを使い，時間とともにどう変化しているのかを分析するのが基本になります．

データは，各都道府県別と全国の日毎の新規陽性者数（自治体で感染を把握した人数）と死者数のデータです．時系列の分析では，全国の系列のデータを分析します．各都道府県の系列も同様の方法で分析できますが，人口の少ない都道府県の日毎のデータは，小さなクラスターや 1 人の陽性，死亡の影響が大きく出るので，注意が必要です．

あとで示すように，曜日により陽性者数，死者数が異なり，その影響を除去する加工（単純移動平均）をしたデータを使った方がよいとされています．そのとき，加工前のデータの系列を「原系列」と呼ぶことにします．

11.2.1　原系列の分析表の作成

原系列の分析用のワークシート「時系列分析」を作成します．そこに，図 21 のような原系列の表を作成し，「=」を使ってデータを取得します．表は，サンプル数（行数）の多い新規陽性者に合わせ，死者数のデータが無いセルは空値（未入力の状態，「””」）にします．取得するデータは，つぎのようになります．

行	取得元のワークシート	取得元のセル範囲	取得先セル範囲
日付	都道府県別新規陽性者数	A2:A1190	A4:A1192
新規陽性者数	都道府県別新規陽性者数	B2:B1190	B4:B1192
死者数	都道府県別死者数	B2:B1076	C118:C1192

	A	B	C	D	E
1	都道府県	全国			
2		原系列			
3	日付	新規陽性者数	死者数		
4	2020/1/16	1			
5	2020/1/17	0			
6	2020/1/18	0			
7	2020/1/19	0			
116	2020/5/7	107			
117	2020/5/8	88			
118	2020/5/9	108	0		
119	2020/5/10	66	8		
120	2020/5/11	58	22		

一部省略

図 21　原系列表

(1) 図 21 のように表頭 (A1:C3) を作成します．「原系列」のセル (B2,C2）は，セルの結合をしています．

(2) 日付と新規陽性者を「=」とスピルを使って取得

A4: =都道府県別新規陽性者数!A2:A1190

B4: =都道府県別新規陽性者数!B2:B1190

(3) 死者数を「=」とスピルを使って取得．死者数は，2020/5/9〜であることに注意

C118: =都道府県別死者数!B2:B1076

動画: 時系列分析，準備

11.2.2 原系列のグラフの作成

時系列データのグラフ化は，時間間隔が等間隔の時（この Covid-19 のデータは日毎 (1 日間隔) のため等間隔）は折れ線グラフまたは散布図でグラフ化し，等間隔ではないとき（例：直近のデータは 1 年間隔，古いデータは 5 年間隔などの統計表）は散布図を使います．

横軸 (X 軸) の値が数値 散布図または折れ線を利用．「2023/3/10」などの日付形式は，通常数値データで，ある基準日から何日経過したかの値で表されているので散布図を利用可能．

横軸 (X 軸) の値が文字列 折れ線を利用．ただし，誤解を与えるグラフを作成しないように，間隔が等間隔であることを確認する．例えば，日毎，月ごと，年度（年）毎の値であるか確認する．

(1) グラフ化する範囲（日付，新規陽性者数，死者数），A3:C1192 を範囲指定します．

(2) メニュー 挿入 → グラフ → 折れ線/面グラフ → 折れ線 (左上のアイコン)

図 22 上は，上記手順で作成した折れ線グラフです．新規陽性者数と死者数は，桁違いな差異があるため，少ない方の死者数

の変化を見ることができません．そこで，図 22 下のようにフィルター機能（グラフ内をクリックすると表示されるボタンから，フィルターのボタン を押し，系列を死者数のみにし，適用をクリックします）で死者数の系列のみします．

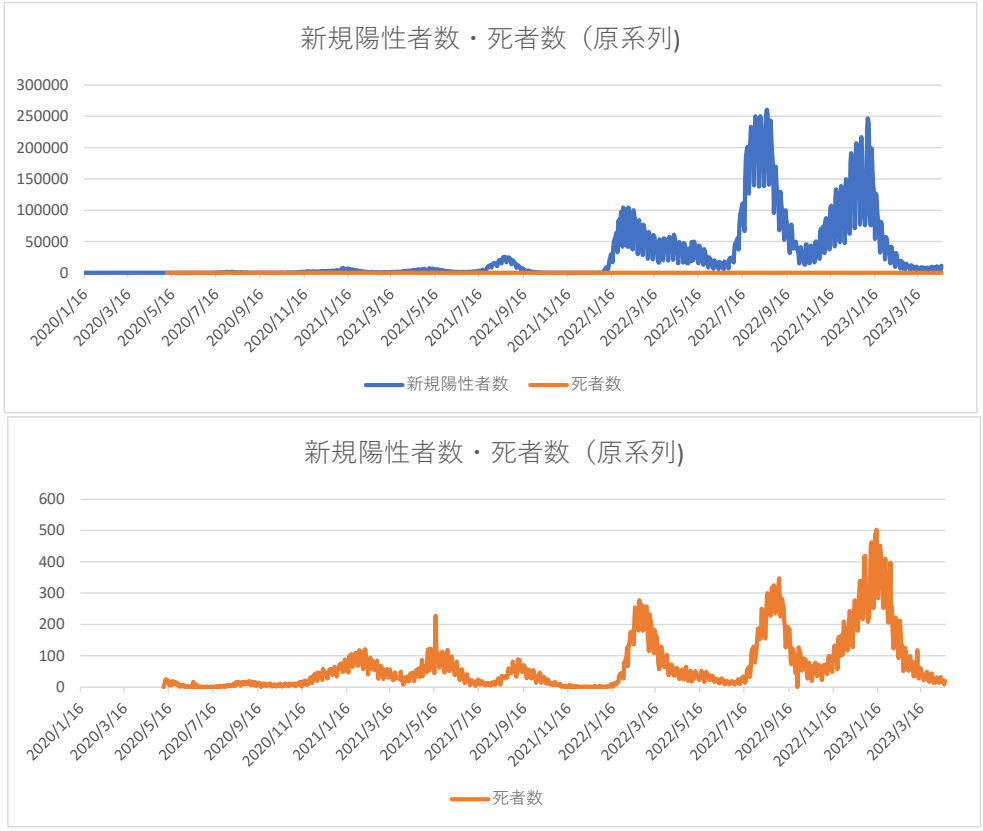


図 22　原系列の折れ線グラフ（下：フィルターを使い，死者数のみ表示）

動画: 時系列分析，原系列のグラフ化

図 22 のグラフには，2 つの問題点があります．

- 1 つのグラフで，新規陽性者数と死者数の変化を観察できない．これは，2 軸グラフを利用する（11.2.3 節で学修），片対数グラフにする（11.4 節で学修），これら 2 つの方法を併用することで対応します．
- 大きな周期的な変動（グラフでは，山とそれらの あいだ の谷）と細かい振幅（おおよそ 7 日くらいの周期）があります．細かい 7 日の周期的な変動は，曜日による医療機関・自治体の対応件数の差異などによって引き起こされると考えられ，分析には含めないことにします．そこで，単純移動平均 (11.3 節で学修) という方法で 7 日周期の周期的な変動を除去することで対応します．

11.2.3　2 軸グラフの作成

2 軸グラフでは，図 23 のように，新規陽性者数を左の Y 軸, 死者数を右の Y 軸（第 2 軸）にします．

- 図 22 の上の状態のグラフを右クリック
- グラフの種類の変更 → グラフの種類の変更 → 組み合わせ
- 新規陽性者数 を 折れ線 のなかの 折れ線 を選択　（第 2 軸には，チェックを入れない）
- 死者数 を 折れ線 のなかの 折れ線 を選択し，第 2 軸のチェックボオクスにチェックを入れる．
- OK をクリック

図 24 のグラフが表示されます．

動画: 時系列分析，原系列のグラフ化 2 軸グラフ

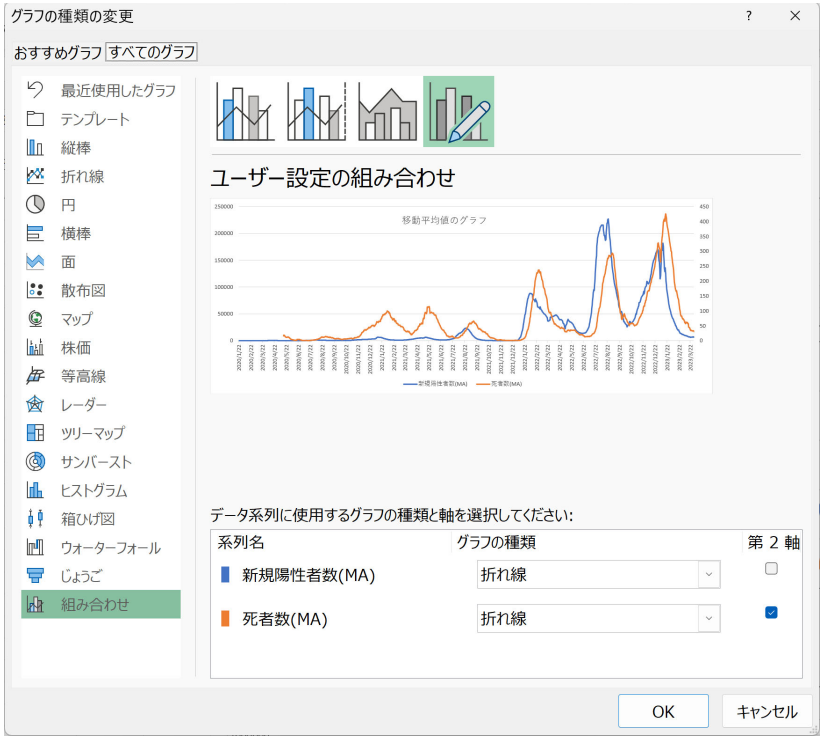


図 23　原系列の 2 軸グラフの設定

11.3　単純移動平均

新規陽性者数，死者数は，曜日による周期的な変動があると述べました．このよう場合，前？ 日の平均値で分析することはよく行われ，？ 日単純移動平均と呼ばれています．このような分析は，株価の推移や外国為替相場の推移などによく用い

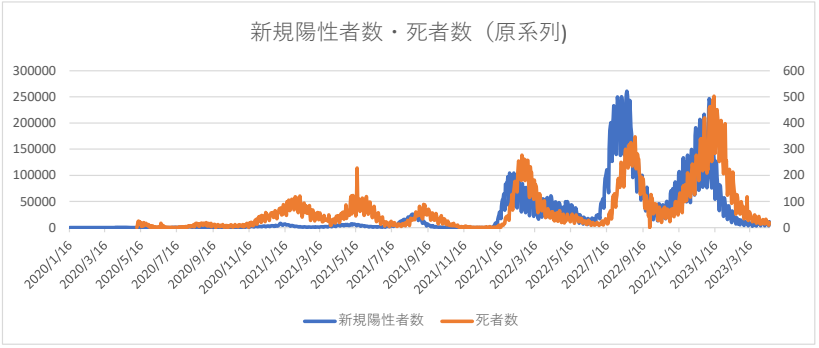


図 24　原系列の 2 軸グラフ（左軸：新規陽性者数，右軸：死者数）

られます.

そこで，1 週間の平均値を使う方法が考えられます．7 日単純移動平均は，対象の日を含めて 7 日の平均値（6 日前～当日までの 7 日の平均値）を用いる方法です．曜日で考えると，ある火曜日の 7 日単純移動平均値は，前の週の水曜日から，その火曜日までの 7 日の平均値です．3/10（3 月 10 日）の単純移動平均値は、次式で求めます.

$$\boxed{3/10\text{ の単純移動平均値}} = \frac{\boxed{3/10\text{ の値}} + \boxed{3/9\text{ の値}} + \boxed{3/8\text{ の値}} + \boxed{3/7\text{ の値}} + \boxed{3/6\text{ の値}} + \boxed{3/5\text{ の値}} + \boxed{3/4\text{ の値}}}{7}$$

$$\boxed{3/11\text{ の単純移動平均値}} = \frac{\boxed{3/11\text{ の値}} + \boxed{3/10\text{ の値}} + \boxed{3/9\text{ の値}} + \boxed{3/8\text{ の値}} + \boxed{3/7\text{ の値}} + \boxed{3/6\text{ の値}} + \boxed{3/5\text{ の値}}}{7}$$

3/10 の移動平均値から 3/11 への移動平均値の差異は，3/11 の値を計算に含め，3/4 の値を計算に含めないようにしたものです.

11.3.1　単純移動平均の計算表

図 25 のように新規陽性者数，死者数の 7 日単純移動平均を求めます．表頭は，新規陽性者数 (MA), 死者数 (MA) としました．MA は，移動平均 (Moving Average) の略です．

(1) 単純移動平均値は，7 日分のデータが揃わないと計算できないので，新規陽性者数は，2020/1/22 からになります．
D10 に，2020/1/16〜2020/1/22 の平均値を求めます．前 7 日の平均なので，B4:B10(7 個のセル) の平均値を求めます．
D10: =AVERAGE(B4:B10)
図では，セルの書式設定で少数点以下 2 桁に設定しています．

(2) D10 の計算式を複写します．
複写元: D10　複写先: D10:D1192

(3) 死者数は，2020/5/9〜ですので，2020/5/15 から計算できます．
E124: =AVERAGE(C118:C124)
複写元: E124　複写先: E124:E1192

動画: 時系列分析，移動平均の計算式

11.3.2　単純移動平均のグラフ

次のように作成します．

(1) グラフ化する範囲（日付，新規陽性者数，死者数），A3:A1192 と D3:E1192 を範囲指定します．2 つめの範囲は，Ctrl

	A	B	C	D	E	F	G
1	都道府県	全国					
2		原系列		7日単純移動平均			
3	日付	新規陽性者数	死者数	新規陽性者数(MA)	死者数(MA)		
4	2020/1/16	1					
5	2020/1/17	0					
6	2020/1/18	0					
7	2020/1/19	0					
8	2020/1/20	0					
9	2020/1/21	0					
10	2020/1/22	0		0.14			
11	2020/1/23	0		0.00			
123	2020/5/14	99	23	80.14			
124	2020/5/15	55	15	75.43	16		
125	2020/5/16	56	19	68.00	18.71429		
126	2020/5/17	29	5	62.71	18.28571		
127	2020/5/18	30	14	58.71	17.14286		

一部省略

図 25　単純移動平均値の計算表

キーを押しながら，マウスで範囲指定します．

(2) メニュー 挿入 → グラフ → 折れ線/面グラフ → 折れ線 (左上のアイコン)

(3) グラフを右クリックし， グラフの種類の変更 → グラフの種類の変更 → 組み合わせ

(4) 新規陽性者数 (MA) を 折れ線 のなかの 折れ線 を選択　（第 2 軸には，チェックを入れない）

(5) 死者数 (MA) を 折れ線 のなかの 折れ線 を選択し，第 2 軸のチェックボオクスにチェックを入れる．

(6) OK をクリック

動画: 時系列分析，移動平均の 2 軸グラフ

単純移動平均のグラフのグラフが表示されますので，X 軸を分析対象の範囲は変更できます (図 26).

(1) グラフの日付の部分を右クリック → 軸の書式指定

(2) 境界値の最小値と最大値を変更します.

例では，新規陽性者の移動平均値がある日 (2020/1/16) から分析の最終日 (2023/3/31) にしていますが，ある特定の期間を分析するのも可能です.

動画: 時系列分析，グラフの表示期間の変更

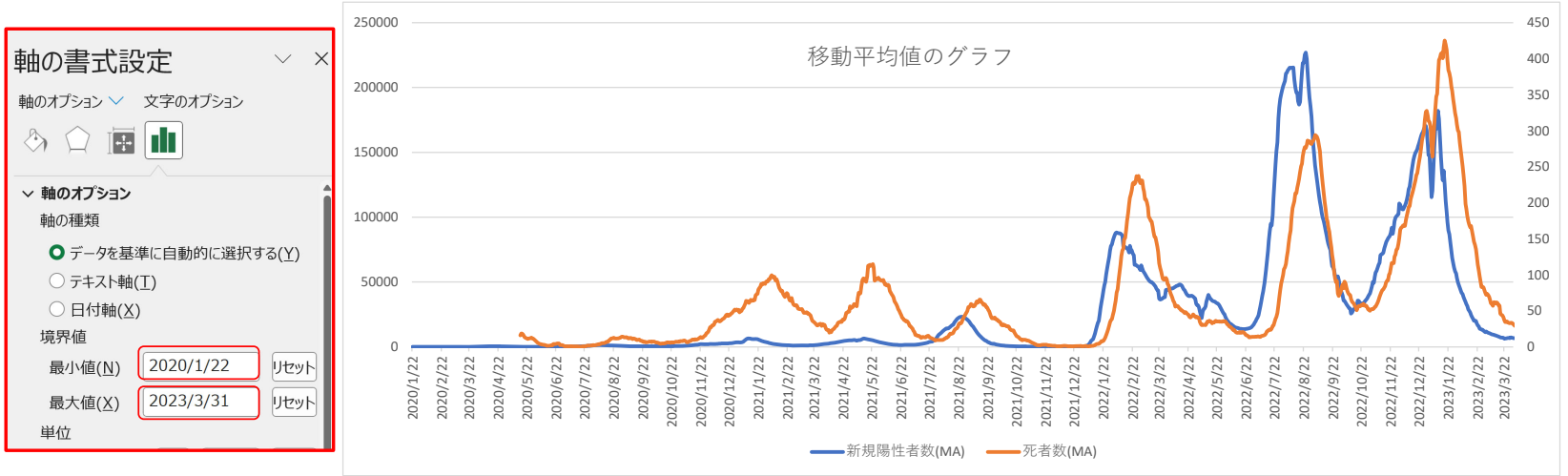


図 26　単純移動平均値のグラフ

図 26 では，週毎の細かい変動はほぼ無くなっており，大きな周期での変化（一般には，「第何波」とよばれています）をみることができるようになっていると思います.

11.4　片対数グラフ

自然現象や社会現象の時系列の変化は，一定期間の変化が何倍に変化していくものが多数あります．感染症の感染者数（新規陽性者数）や死者数は，（前週などの）何倍になるのかで議論されており，倍数で変化しているデータです．このように一定期間に倍数で変化するデータは，縦軸の数値を対数で表示すると変化の動向が見やすくなります．具体的には，図 26 のように，普通の目盛（線型間隔の目盛）ではなく，1 つ目盛が増えると 10 倍（10 倍以外の値にすることもあります）になるように変更した対数目盛にします．このように X 軸と Y 軸のうち，片方の Y 軸のみ対数目盛にしたものは片対数グラフと呼ばれています．

11.5　片対数グラフの作成（2 軸グラフ）

(1) 図 26 をコピーします．図を右クリックし，コピー を選び，適当な場所で，貼り付けオプション で，元の書式を保持 を選びます．

(2) コピー先の左側の軸を右クリックし，軸の書式設定 を選びます．

(3) 軸のオプション の中の 対数目盛を表示する にチェックを入れます．
基数の欄は 10 のままにしておきます．この基数を変更すると何倍になったら 1 目盛増えのかを調整できます．

(4)「負の数値またはゼロは，対数グラフに正しくプロットされません．対数目盛では，正の数値だけが有効です」というメッセージボックが表示されることがあります．OK をクリックしてください．
0 や負の対数は，（実数の範囲では）存在しないからです．0 に近い正の実数の対数は，負の無限大に近い値になります．

(5) 右側の軸も同様に，対数目盛にします．

(6) 図 27 のようなグラフが作成されます．

動画: 時系列分析，2 軸片対数グラフ

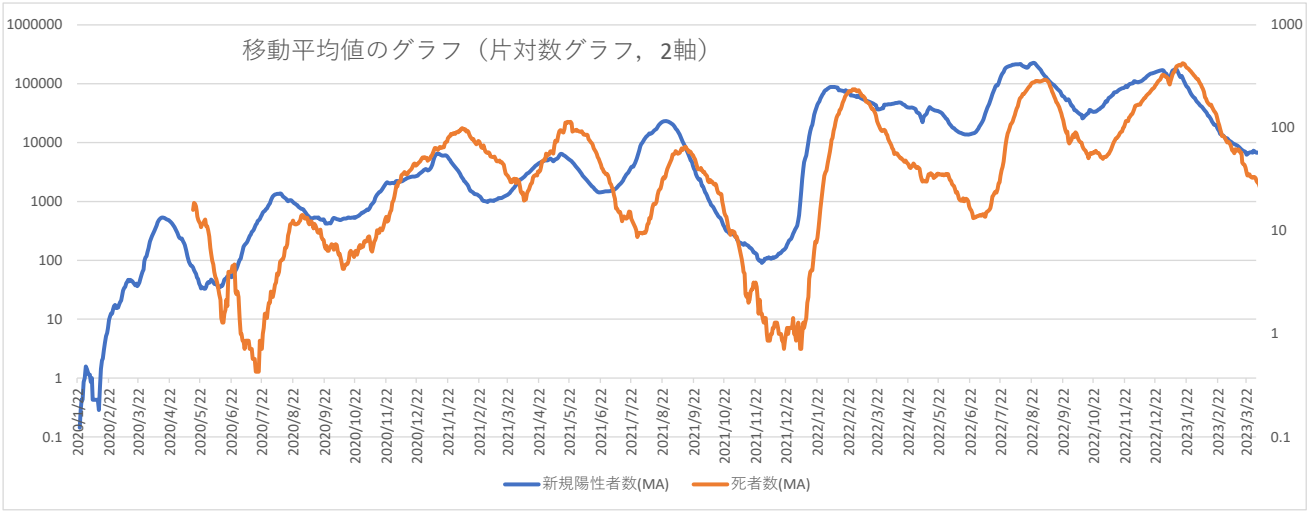


図 27　片対数グラフ (2 軸)

図 27 をみると，それぞれの山と谷がはっきり見えてきたと思います．

11.6　片対数グラフの作成（1 軸グラフ）

片対数グラフにすることにより，図 28 のように 2 軸にしなくても値が小さい死者数の変化を観察てきます．練習のため作成しましょう（分からない場合は動画を参照してください）

動画: 時系列分析，片対数グラフ

図 28 では 1 軸の対数目盛にすると新規陽性者数と死者数は同じ軸の値なので比較できます．通常の目盛の場合その差異

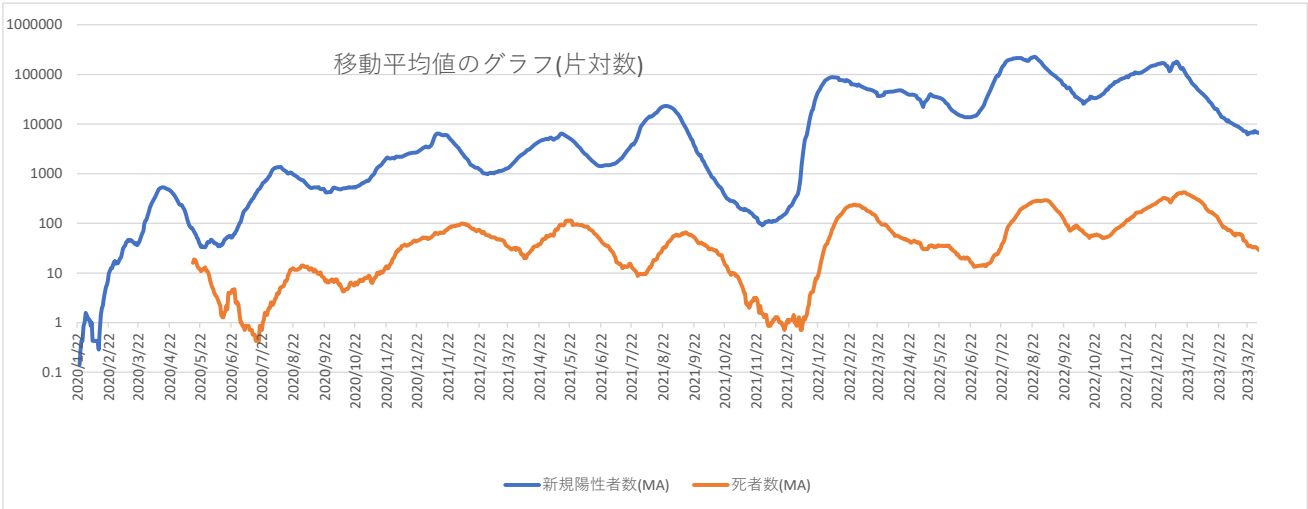


図 28　片対数グラフ (1 軸)

は，「新規陽性者数 − 死者数」の値ですが，対数目盛の場合は何倍か「新規陽性者数 / 死者数」を表しています．例えば，2020/8/11 付近では，2 つの値の差異は約 2 目盛あるので，$10^2 = 100$ 倍の差異があることを示しています．2022/7/22 付近では，3.5 目盛くらい離れているので，3 目盛 1000 倍と 4 目盛 10000 倍の間で，計算上は約 3000 倍差異があることが分かります．

このグラフから，2 つの線の差異は，2021/12 くらいまでは 2 目盛，2022/1 からは差異が拡大しており，3 目盛くらいになっていることが分かります．

11.7　相関係数を利用して遅れを推定 (参考)

図 28 は，全国の感染者数と死者数推移の移動平均値を対数目盛のグラフで表したものでした．グラフをみると，感染者数と死者数の形状が似ており，また，感染者数が先行していて，死者数は 2~3 週間遅れて（遅行して）推移しているのが分かります．そこで，遅れが何日であるのか，相関係数を使って求めてみたいと思います．

授業時間の都合で本節は，説明を省略することがあります．その場合，遅れ日数を 16 として分析をします．

11.7.1　相関係数を使って遅れ日数の推定

遅れの日数を X(非負の整数 (0 以上の整数)) とします．相関係数を求める関数は，CORREL で次のようになります．

= CORREL(系列 1 のセル範囲, 系列 2 のセル範囲)

ただし，系列 1 のセル範囲と系列 2 のセル範囲の大きさは同じであることが必要です．系列 1 は，新規陽性者数 (MA) で，2020/5/15～2023/3/14 の値（セル範囲では，D124:D1158 の 1035 個のセル）とします．系列 2 は，死者数 (MA) で，次のように遅れ日数の開始日から うしろにずらしていき，相関係数が最大になる遅れ日数を求めます．

遅れ日数 X	利用する死者数の日付	MA の利用するセル範囲	セルの個数
0	2020/5/15～2023/3/31	E124:E1158	1035 個
1	2020/5/16～2023/4/1	E125:E1159	1035 個
2	2020/5/17～2023/4/2	E126:E1160	1035 個
:	:	:	:

あるセルから，相対的な位置でセル範囲を指定する関数には，OFFSET 関数があります．OFFSET 関数は次のように使います．

= OFFSET(基準, 行数, 列数, 高さ, 幅)

とします.

基準　どのセルからの相対位置であるかのセル (1 つ）を指定します．この例題の場合，もとの開始位置 (死者数の開始セル) の E124 とします.

行数, 列数　基準からどの程度，右下かを指定します．行数分下，列数分右のセル範囲を返します．この例題の場合，行数は，遅れ日数 (セル E1) 分したのセル，列数は同じ列なので 0 とします.

高さ, 幅　返すセル範囲の行数（高さ）と列数（幅）を指定します．この例題の場合，高さは，系列 1 と同じ 1035 にし，列数は 1 にします.

したがって，系列 2 のセル範囲は,

OFFSET(E124,E1,0,1035,1)

となりります.

図 29 のように，J 列に遅れ日数を指定し，K 列にその相関係数を求めます.

遅れに日数を設定します.

J4:	0

CORREL を使って，D124:D1158 と OFFSET 関数で指定した X 日の遅れの死者数との相関係数を求めます.

K4:	=CORREL($D$124:$D$1158,OFFSET(E$124,J4,0,1035,1))

とします．この式は下に複写して利用します．D124:D1158 と E124 は下に複写しても行番号は変化させないので，絶対参

	A	J	K
1	都道府県	相関係数最大値	0.875600259
2		最大値の遅れ日数	16
3	日付	遅れ日数	相関係数
4	2020/1/16	0	0.73415204
5	2020/1/17	1	0.750596084
6	2020/1/18	2	0.765782447
7	2020/1/19	3	0.779999371
8	2020/1/20	4	0.793248812
9	2020/1/21	5	0.805580441
10	2020/1/22	6	0.817014874

図 29　遅れ日数の推定

照にします (J4 の遅れ日数は変化させるので，絶対参照にはしません)．

J4 の遅れ日数を 0〜30 の範囲で変化させ，相関係数が最大となる値を求めることもできます．変化させる回数も多いので，遅れに日数を 0 から 30 まで 1 ずつ変化させ，最大値を求め，それらから最大となる遅れに日数を求めます．

J5:　=J4 + 1

複写元: J5 , 複写先: J6:J34

複写元: K4 , 複写先: K5:K34

次に，最大の相関係数の値とそれに対応する遅れ日数を求めます．

K1: =MAX(K4:K34)

K2: =MATCH(K1,K4:K34,0)-1

K1 は，各遅れ日数の相関係数の最大値です．MATCH 関数は，ある値が検索範囲の何行目 (1 行目を 1 とする) に一致するのかを求めるものです．

MATCH(検索値，検索範囲,0)

と記述し，0 は完全一致するものを検索します．MATCH 関数はセル範囲の一番上の値と一致するとき 1 を返すので，一番上は遅れ日数は 0 なので，1 を引いて遅れ日数にします．

動画: 時系列分析，相関係数を使い遅れ日数を推定（旧）

遅れ日数が 16 のとき，相関係数が 0.8756 となり最大となったと思います．この 16 日の遅れを固定して分析します．

11.7.2 系列，X 日遅れ死者数 (MA) の作成とグラフ化

図 30 のように X 日遅れ死者数 (MA) を作成します．16 日遅れですので, 図 30 のように, 2020/5/15 の 16 日前の 2020/4/29 の「X 日遅れ死者数 (MA)」(F108) は，2020/5/15 の「死者数 (MA)」(E124) の値とし，以下のセルも同様にします．

F108: =E124

複写元: F108 複写先: F108:F1176

動画: 時系列分析，X 日遅れの計算

A3:F1192 を範囲指定して，折れ線グラフを作成し，Y 軸を対数目盛にします．ただし，グラフフィルターで新規陽性者数 (MA)，死者数 (MA)，X 日遅れ死者数 (MA) のみ表示するようにしています．

動画: 時系列分析，X 日遅れ等の片対数グラフ

図 31 のようなグラフになったと思います．新規陽性者数 (MA) の 16 日遅れがほぼ 16 日遅れ死者数 (MA) に対応してい

	A	B	C	D	E	F	G	H	I
1	都道府県	全国					全国(千人)	124885.175	
2		原系列		7日単純移動平均			人口千人当たり　7日移動平均		
3	日付	新規陽性者数	死者数	新規陽性者数（MA）	死者数（MA）	X日遅れの死者数(MA)	新規陽性者数（MA）	死者数（MA）	X日遅れの死者数(MA)
06	2020/4/27	182		356.428571			0.00285405		
07	2020/4/28	279		340.714286			0.00272822		
08	2020/4/29	217		308		16	0.00246627		0.00012812
09	2020/4/30	202		274.142857		18.71429	0.00219516		0.00014985
10	2020/5/1	282		251		18.28571	0.00200985		0.00014642
11	2020/5/2	288		239.142857		17.14286	0.00191490		0.00013727
12	2020/5/3	195		235		14.71429	0.00188173		0.00011782
13	2020/5/4	181		234.857143		12.85714	0.00188058		0.00010295
14	2020/5/5	119		212		12.28571	0.00169756		0.00009838
15	2020/5/6	104		195.857143		11.85714	0.00156830		0.00009494
16	2020/5/7	107		182.285714		10.85714	0.00145963		0.00008694
17	2020/5/8	88		154.571429	16日遅れ	11.57143	0.00123771		0.00009266
18	2020/5/9	108	0	128.857143		11.85714	0.00103180		0.00009494
19	2020/5/10	66	8	110.428571		12.42857	0.00088424		0.00009952
20	2020/5/11	58	22	92.8571429		12.85714	0.00074354		0.00010295
21	2020/5/12	87	25	88.2857143		11.14286	0.00070694		0.00008922
22	2020/5/13	55	19	81.2857143		11.14286	0.00065088		0.00008922
23	2020/5/14	99	23	80.1428571		10.14286	0.00064173		0.00008122
24	2020/5/15	55	15	75.4285714	16	8.857143	0.00060398	0.00012812	0.00007092
25	2020/5/16	56	19	68	18.71429	6.857143	0.00054450	0.00014985	0.00005491

図 30　X 日遅れの死亡者数の表

ることが分かると思います.

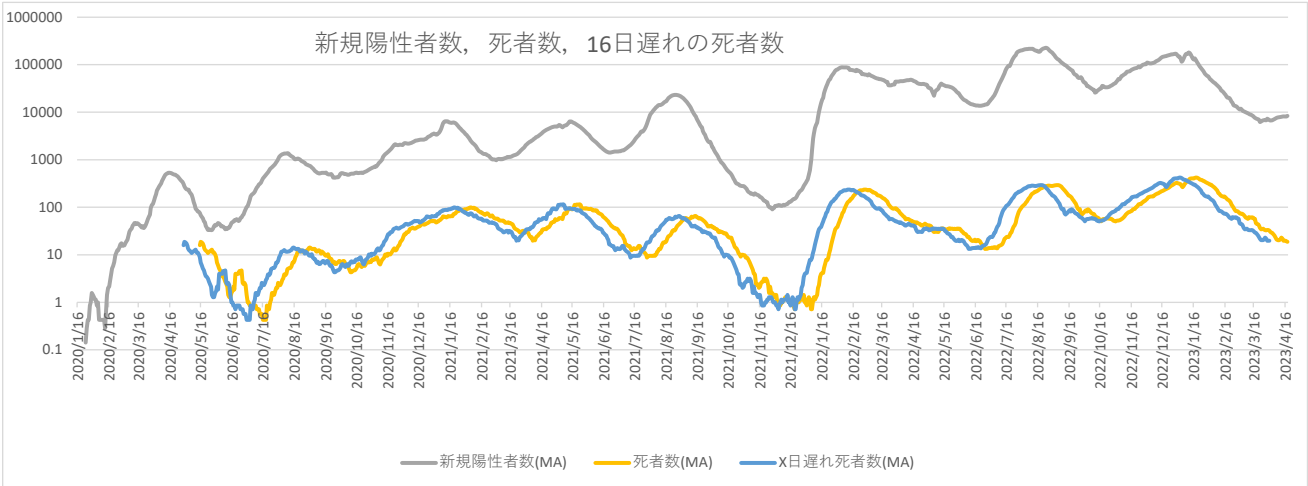


図 31　X 日遅れ死者数 (MA) のグラフ化

11.8　変異株別の致死率の計算

変異株により，致死率が異なると言われています．本節では，それを確認するために変異株毎の感染者数，死亡者数を求め，それらから致死率を求めます．[*1]．

図 32 は，新規陽性者と死者数の移動平均値のグラフと，波が第何波（赤〇）かと主に流行した COVID-19 の変異株を示したものです．波の番号は，新規陽性者の波で示しています．変異株の種類は，赤い矢印で範囲を示しています．

[*1] 2024 年度までは，日毎の致死率を計算していましたが，感染の波により致死率が，一時的に上がったり下がったりするので，2025 年度からは，COVID-19 の変異株毎の致死率を計算することにしました．

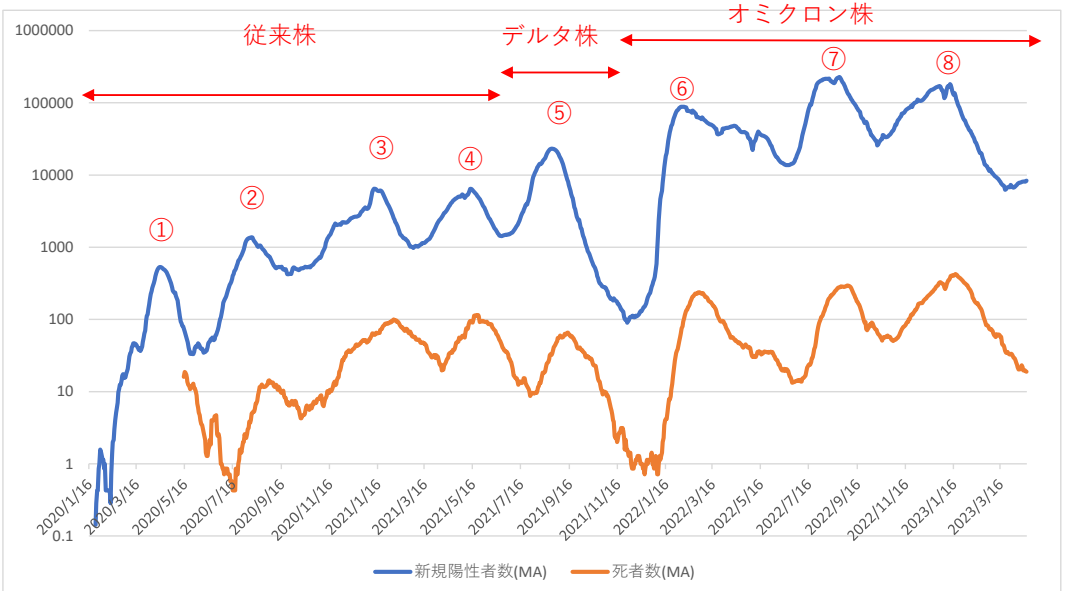


図 32　新規陽性者と死者数の移動平均値と第何波，変異株

表 1 は，波により A,B,C の 3 期間に分け，新規陽性者数と死者数それぞれの範囲を求めました．新規陽性者数と死者数には，16 日（前節で求めた値）の差を付けています（16 行の差）．

11.8.1　練習問題

図 33 は，各期間について，新規陽性者数と死者数の合計を，それぞれの範囲を使って求め，致死率を求める計算表です．たとえば，期間 A の計算式は，次のようになります．新規陽性者数の範囲は，表 1 より 118 行目から 535 行目までなので，

=SUM(B118:B535)

表 1　波と変異株による期間の分割

期間名	期間	日数	波	（変異）株	新規陽性者範囲	死者数範囲
全期間	2020/5/9～2023/3/31	1057	第 1 波～第 8 波	–	118～1174	134～1190
期間 A	2020/5/9 ～2021/6/30	418	第 1 波～第 4 波	従来株	118～535	134～551
期間 B	2021/7/1～2021/12/31	184	第 5 波	デルタ株	536～719	552～735
期間 C	2022/1/1～2023/3/31	455	第 6 波～第 8 波	オミクロン株	720～1174	736～1190

期間名	波	株	新規陽性者数	死者数	致死率
全期間	第1波～第8波	–			
期間 A	第1波～第4波	従来株			
期間 B	第5波	デルタ株			
期間 C	第6波～第8波	オミクロン株			

図 33　各期間の新規陽性者数と死者数，致死率

で求めることができます．同様に，死者数は表 1 より C 列の 134 行目から 551 行目までなので，=SUM(C134:C551) という計算式で，致死率は死者数÷新規陽性者数で計算します．

図 33 の計算表を作成し，計算式で空欄の値を求めなさい．

12　付録：e-Stat から家計調査のデータをダウンロードする方法（参考）

データサイエンス入門で学習した e-Stat の家計調査を利用して，2021 年の都道府県別，2 人以上の世帯のデータを元に，消費支出額，5 品目（例題では米，パン，ケーキ，せんべい，チョコレート）の平均支出額について分析してみます．年度を変えて分析をすることも可能ですし，品目を変更して練習することも可能です．

(1) https://www.e-stat.go.jp/ にアクセス
(2) 検索等を用いて「家計調査」のデータベースを選択
(3) 家計収支編，二人以上の世帯，年次　を選択
(4) 品目分類の中の　「010　品目分類（2020 年改定）（総数：金額）」の DB を選択
(5) 表示項目選択の品目分類：「消費支出額」と 5 品目（例題では「米」,「パン」,「ケーキ」,「せんべい」,「チョコレート」) を選択
(6) 表示項目選択の世帯区分：「二人以上の世帯」のみ選択
(7) 表示項目選択の地域区分：全国と全県庁所在地，政令指定都市を選択.
(8) 表示項目選択の時間軸：ある年 (一年) を選択（例題：2021 年のみ選択）
(9) レイアウト設定
    - 列：品目分類
    - 行：地域区分
    - ページ上部：表章項目，世帯区分，時間軸
(10) ダウンロード

動画: 家計調査，データの取得

	A	B	C	D	E	F	G	H	I
1	統計名：	家計調査 家計収支編 二人以上の世帯							
2	表番号：	010							
3	表題：	[品目分類] 品目分類（2020年改定）（総数：金額）							
4	実施年月：	-	-						
5	市区町村時	-							
6	表章項目：	金額							
7	世帯区分	二人以上の世帯（2000年～）							
8	時間軸（年	2021年							
9									
10	地域区分	/品目分類	費支出【円	1.1 米【円】	2 パン【円	ケーキ【円	せんべい【円	コレート【円】	
11	全国		3,348,287	21,862	31,353	7,716	5,719	6,664	
12	01100 札幌市		3,220,749	27,371	27,609	8,339	4,427	6,731	
13	02201 青森市		2,941,401	20,934	25,662	6,563	5,955	4,861	
14	03201 盛岡市		3,274,937	22,927	28,638	7,984	6,249	5,678	
15	04100 仙台市		3,410,054	21,594	30,170	8,200	6,141	6,386	
16	05201 秋田市		2,973,311	21,126	23,775	6,337	4,623	5,837	

図 34　取得した家計調査ファイル (ワークシート名:1)

図 34 は，取得した家計調査ファイルです．図 1 は，図 34 のデータから，分析しやすいように整理したものです．

- 消費支出【円】 1.1.1 米【円】 1.1.2 パン【円】 344 ケーキ【円】 350 せんべい【円】 352 チョコレート【円】 などのコードや「【円】」は手作業で除去します.
- 市名の前にコードが付いています．MID 関数を用いてコードや空白を除去します.
  - 新しいワークシートを作成し，ワークシート名を入力します（例えば,「dataset」).
  - 「全国」は，そのまま全国にします.
  - A4: =REPLACE('1'!A12,1,6,"")
    REPLACE 関数は，最初の引数 ('1'!A12) の第 2 引数（1 文字目）から第 3 引数（6 文字目）までを第 4 引数 ("" , 空文字）に置き換えます.
    「'1'!A12」は，シート「1」の A12 を表します．シート「1」をクリックして，A12 をクリックします.
  - 複写元: A4 , 複写先: A5:A55
- 支出金額などは，データサイエンス入門で指摘したように文字列になっていることがあります．そこで，VALUE 関数を用いて文字列を数値に変換します．文字列を数値に直しながら，計算式で値を転記します.
  - B4: =VALUE('1'!C11)
  - 複写元: B4，複写先: B4:G55

動画: 家計調査，データの整理

13　付録: きりの良い階級にする（参考）

きりのよい数を境界に度数分布表（ヒストグラム）を作成することが多く，その方が人間にとって扱いやすくなります.
これを実現するにはテクニックが必要です（「ヒストグラムの作成 with Excel 4/4」でのテクニックを参考にしました).
ヒストグラムの縦棒は Excel では「ビン」と表現されています．最初（一番左）の階級を「ビンのアンダーフロー」，最後

（一番右）の階級を「ビンのオーバーフロー」に割り当てて，あいだの部分は等幅のビンで描画させます．「ビンのアンダーフロー」と「ビンのオーバーフロー」の境界値と，あいだの部分のビンの幅は指定できるので，

- 「ビンのアンダーフロー」には，1 つ目と 2 つ目の境界の値
- 「ビンのオーバーフロー」には，最後の 1 つ前の階級と最後の階級の境界の値

とします（図 35）．

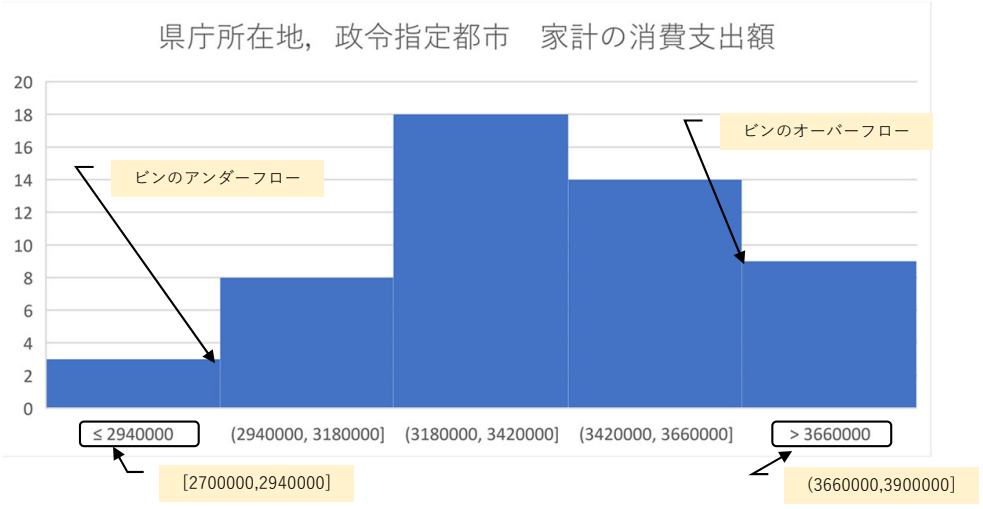


図 35　きりの良い階級のヒストグラム

最初の階級の下限と階級幅を計算

- 区間数は，自動で作成したヒストグラムの個数のままのほうが良いようです．この場合，5

	A	B	C	D	E	F	G
56							
57		平均	3,369,139			2,700,000	←最初の階級下限
58		中央値	3,398,723			2,940,000	←H57＋幅
59		標準偏差	277,456			3,180,000	←H58＋幅
60		最大値	3,873,418			3,420,000	
61		最小値	2,708,442			3,660,000	
62		階級数	5			3,900,000	←最後の階級上限
63		幅	232,995				
64		度数分布表設定					
65		最初の階級下限	2,700,000	←最小値よりやや小さい　きりの良い値を手入力			
66		幅	240,000	←B63よりやや大きい　きりの良い値を手入力			
67		階級数	5	←B62と同じ値			
68		最後の階級上限	3,900,000	←計算式(最初の階級下限＋幅＊階級数)			
69			↑最大値よりやや大きいことを確認				

図 36　きりの良い階級を求める

- 最大値，最小値，区間数（この場合５で固定），幅を求めます．
  幅は，（最大値 – 最小値）÷ 区間数
- C65 から C68 は，最初の階級の下限と階級幅を与えて，最初の階級の下限と最後の階級の上限の値のあいだに全サンプルの値が入るようにします．
- C65:C67 は，手入力とします．
- C68: =C65+C66*C67
  (最初の階級下限＋幅＊階級数）で，最後の階級の上限を求めます．

- C65 は，最小値 (C61) よりやや小さい きりのよい値とします.
- C66 は，幅 (C63) よりやや大きい きりのよい値とします.
- C68 が，最大値よりやや大きくなるように C65，C66 を調整します.

動画: ヒストグラム，最初の階級の下限と階級幅を計算

最初の階級の下限，階級の境界，最後の階級の上限 (F57:F62) を求める

- F57: =C65
- F58: =F57+$C$66
  1 つ上のセルの値に階級幅を加えていく
- 複写元: F58 , 複写先: F59:F62
  F57:F62 は，階級数 +1 個のセル

動画: ヒストグラム，最初の階級の下限，階級の境界，最後の階級の上限を求める

きりの良い階級のヒストグラムの作成

(1) グラフのヨコ軸を右クリックし，軸の書式設定 → 軸のオプション → 軸のオプション

(2) ビンの幅をチェックし，ビンの幅を階級幅（図 36 の C66 の値，240000）に設定

(3) ビンのオーバーフローにチェックし，最後の 1 つ前の階級と最後の階級の境界の値 (F61 の値，3660000) を入力

(4) ビンのアンダーフローにチェックし，1 つ目と 2 つ目の境界の値 (F58 の値，2940000) を入力

動画: ヒストグラム，きりの良い階級のヒストグラムの作成

完成したヒストグラムで，「≤ 2940000」の部分は，「$[2700000, 2940000]$」と記載した方がより適切ですので，描画ツールで書き直すことも 1 つの手段です.