第 3 章
表計算ソフトウエア：簡単な統計分析

2022 年 8 月 31 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．

本章は，専修大学商学部の高萩栄一郎の著作である．
e-Stat の家計調査，COVID-19 の都道府県別の感染数などのデータについて簡単な統計分析をします．

1　家計調査データから消費支出と 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	H
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．なお，1 行目の 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 B1:D54 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 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．

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 8 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，系列の番号を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．

■index 関数　index 関数は，

INDEX(配列, 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．配列の部分は，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($A$1:$D$5,G2,G3)」の値を表示しています．
- 配列$A$1:$D$5 は，通常固定されるので，絶対参照にします．

- G2 に行番号が記述し，G3 に列番号が記述し，
- 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		配列	$A$1:$D$5		
2	B1	B2	B3	B4		行番号	2		
3	C1	C2	C3	C4		列番号	3		
4	D1	D2	D3	D4		値	B3		
5	E1	E2	E3	E4			↑=INDEX($A$1:$D$5,G2,G3)		
6									

図 2　index 関数

■列番号を指定してその列の値を取り出す　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．

- 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おきます．
- 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)．
- B 列に地域区分を複写します．
- 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．
- C1 の値を元に，Index 関数を使って，dataset から値を取り出します．

	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	

	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)

  - C1: =INDEX(dataset!$B$2:$G$55,A2,$C$1)
  - 配列は絶対参照にします．列番号も配列にします．
  - 行番号（A2）は変化するので，相対参照 ($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

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．

	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，政令市都市を 1 つの単位と見なし，その平均値，中央値，標準偏差，Z 値を求めます (図 5)．標準偏差は，データの平均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）の値と平均値の差異（偏差と言います）の広い意味での平均値で，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少ないことになります．

平均値などを求めるのあたって，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．また，ここで求める各都市の平均と全国の値は少し異なります．これは，全国の値が全世帯の平均（情報入門 1 で学習

	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

した加重平均に近い値）であることからです．

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．

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．

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

平均点の時，Z 値は 0 なので偏差値は 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．

(2) D4: =(C4-$C$57)/$C$59

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．

	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)

(3) E4:　=D4*10+50

(4) 複写元:　D4:E4　　複写先:　D4:E55

動画:Z 値，偏差値の計算

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見ましょう．

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 には自動で図 7 のように，階級（範囲）をいくつか設定し，各階級にいくつのサンプルがあるか (度数) の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.

(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. 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

- [2708442, 2968442] は，2708442 以上,8968442 以下を示しています．グラフから 2708442 以上,8968442 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十分です.

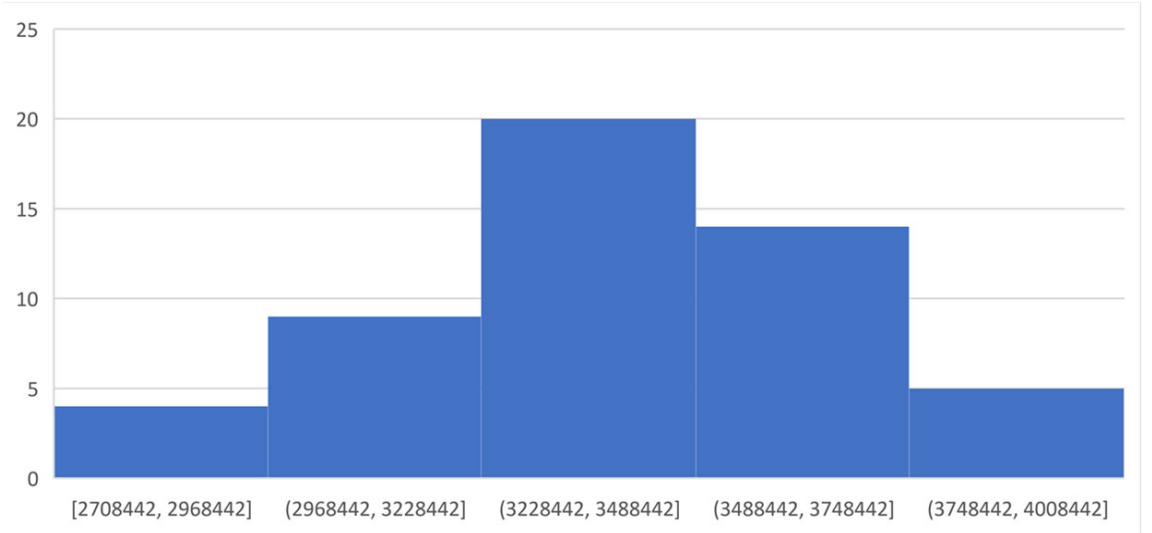


図 7　Z 値，偏差値

- 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．

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

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 9 に載せました．

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．グラフでは，散布図 (XY グラフ) でします．

■**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
(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が見るの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 軸の値が大きいという関係があるとき 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とき −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 F3 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) F3 =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
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．当たり前ですが，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．このような関係の時，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いう傾向はないようです．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．

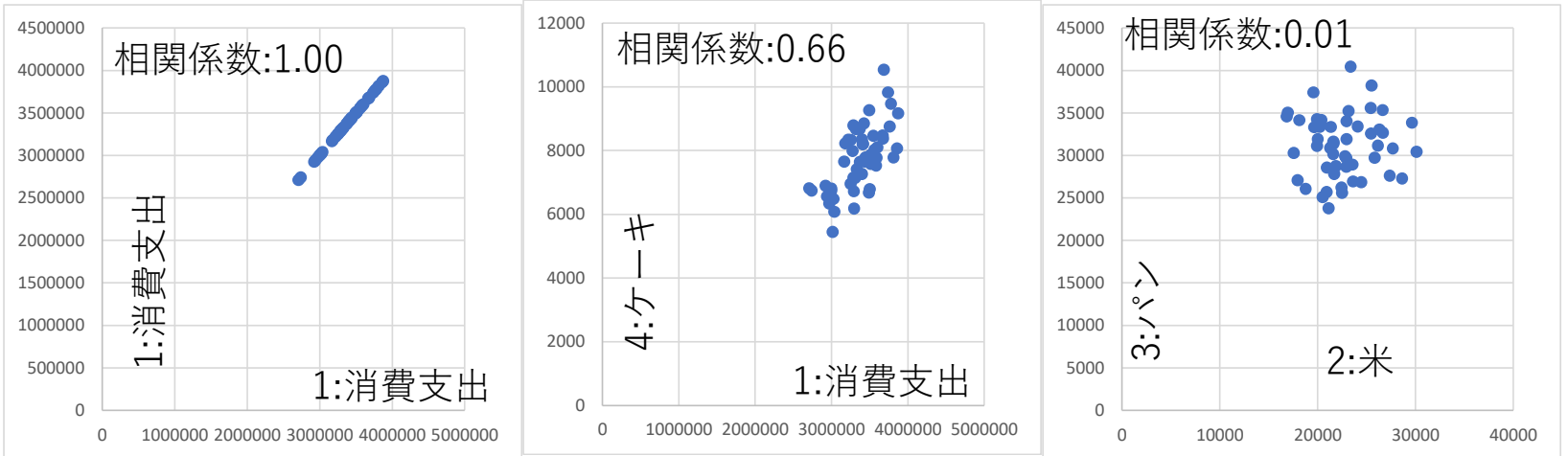


図 9　3 つの散布図と相関係数

この例では出ませんでしたが，散布図の点は，左上から右下への直線上に近い場所のあるとき，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．正の相関関係も負の相関関係も 1 または −1 に近いほど，正または負の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 のような表の形に纏めたものを相関係数行列と呼びます．左上から右下に掛けては，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) の相関係数は同じであるので，左上から右下の 1 の線を対称に値は等しくなります.

	消費支出	米	パン	ケーキ	せんべい	チョコレート
消費支出	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

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　値貼り付けをします．

(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，直感的には理解するため，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してします（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 → 新しい書式ルール

(3)「新しい書式ルール」のウインドウが表示されます（図 12）．

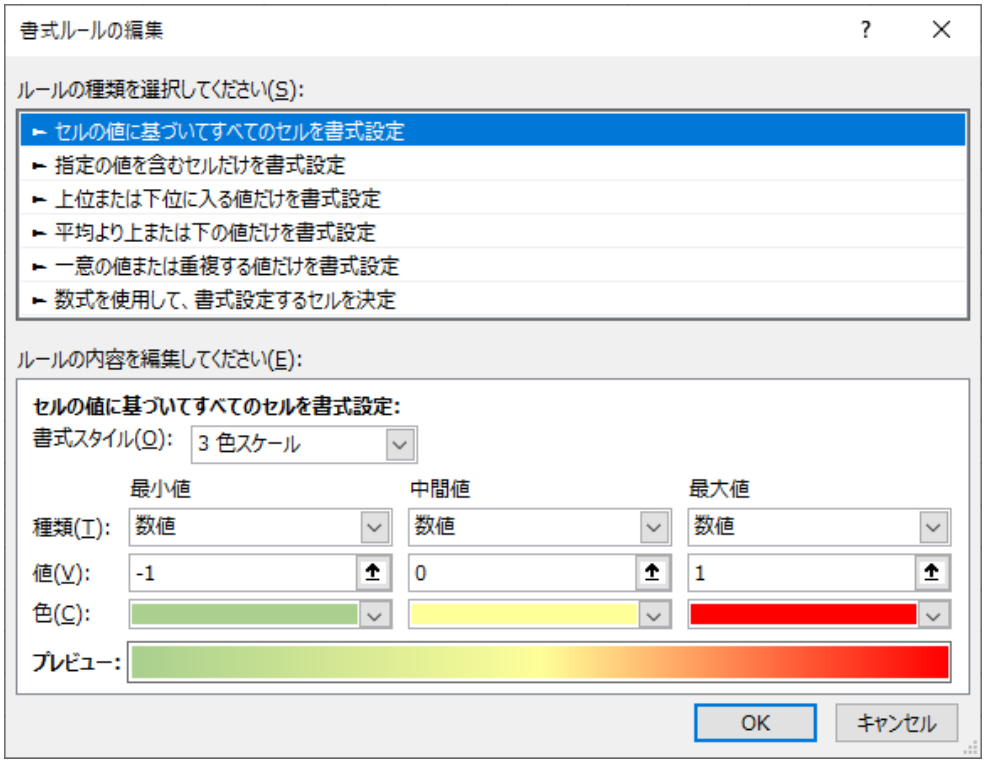


図 12　相関係数行列のヒートマップの設定

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

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) あたら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．

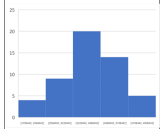
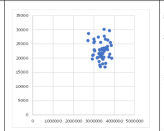
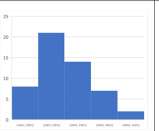
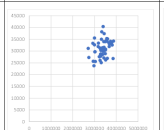
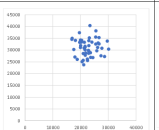
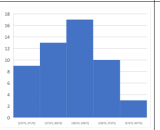
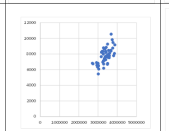
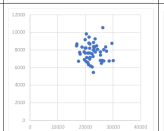
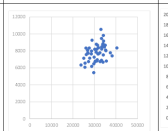
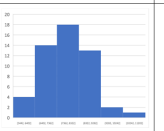
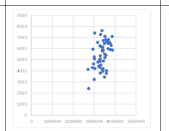
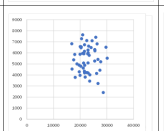
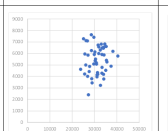
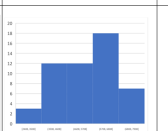
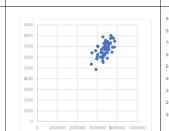
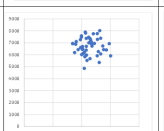
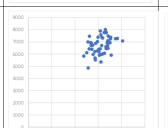
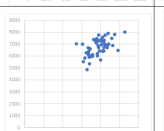
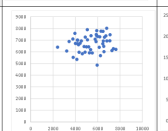
	消費支出	米	パン	ケーキ	せんべい	チョコレート	
消費支出		-0.04	0.30	0.66	0.45	0.61	消費支出
米			0.01	-0.06	-0.06	-0.02	米
パン				0.36	0.08	0.39	パン
ケーキ					0.21	0.54	ケーキ
せんべい						0.18	せんべい
チョコレート							チョコレート
	消費支出	米	パン	ケーキ	せんべい	チョコレート	

図 13　相関図行列

(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.

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，一方の値が多きければもう片方の値が大きな値であるという関係 (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，一方の値が多きければもう片方の値がちいさな値であるという関係 (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 0.66 と 0.61 と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？

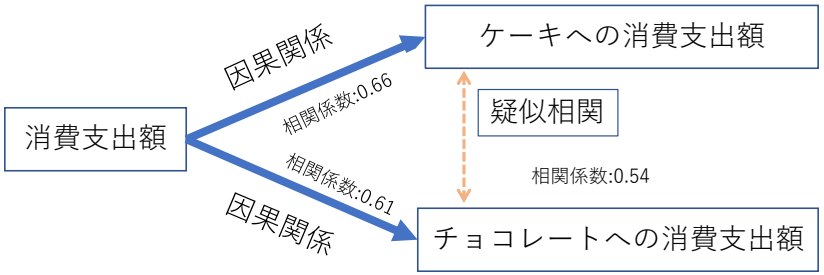


図 14　疑似相関

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あることが起こりました．このように，別の共通の原因で相関があることは疑似相関と呼ばれています．

7　相関係数を利用して遅れを推定 (COVID-19)

7.1　感染者と死者数の推移

図 15 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 2 軸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，死者数は数 10 日遅れて（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．

また，感染者数と死者数の散布図を作成し，どのように変化してきているのかも分析したいと思います．

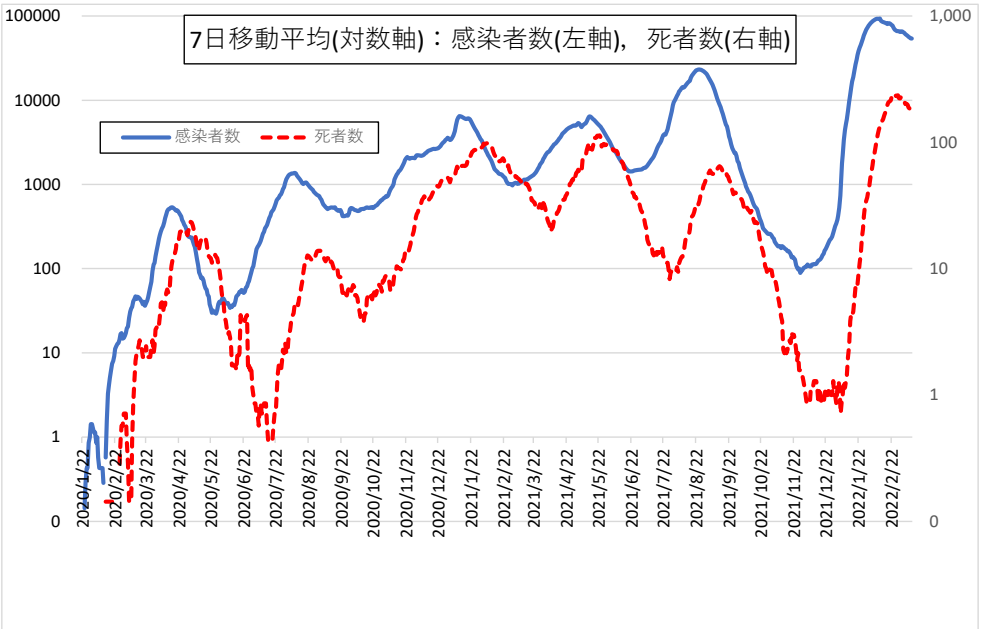


図 15　全国感染者数・死者数推移 (7 日単純移動平均・対数目盛)(情報入門 1 第 5 章 図 21)

7.2　死者数を遅れさせた系列の作成

情報入門 1 で作成した S10_**.xlsx で作成した 7 日移動平均のワークシート (ワークシート名:推移分析) の値を利用し，本分析用のワークシート「遅れの分析」を追加作成し，そこで分析をします．

図 15 をみると，感染者数のピークに数週間遅れて死者数のピークが訪れているのが分かります．遅れを分析するにあたって，死者数を X 日遅れさせた系列を作成します．遅れは最大 30 日くらいでの分析が考えられ，2022/1/20 くらいまで分析

できそうですが，このグラフでは第 6 波は収束していません．そこで，感染者数で第 5 波が収束した（感染者数で底を記録した）2021/11/30 までを分析対象にします．

	A	B	C	D	E	F	G
1	遅れ日数	14			相関係数	0.5183	
2							
3					単純移動平均	単純移動平均	
4	日付(感染者数)	日付(死者数)		日付(感染者数)	感染者数	死者数	
5	2020/1/22	2020/2/5		2020/1/22	0.142857143	0	
6	2020/1/23	2020/2/6		2020/1/23	0	0	
7	2020/1/24	2020/2/7		2020/1/24	0.142857143	0	
8	2020/1/25	2020/2/8		2020/1/25	0.285714286	0	
9	2020/1/26	2020/2/9		2020/1/26	0.428571429	0	
10	2020/1/27	2020/2/10		2020/1/27	0.428571429	0	
11	2020/1/28	2020/2/11		2020/1/28	0.857142857	0	
12	2020/1/29	2020/2/12		2020/1/29	1	0	
13	2020/1/30	2020/2/13		2020/1/30	1.428571429	0.142857143	
14	2020/1/31	2020/2/14		2020/1/31	1.428571429	0.142857143	
15	2020/2/1	2020/2/15		2020/2/1	1.285714286	0.142857143	
16	2020/2/2	2020/2/16		2020/2/2	1.142857143	0.142857143	
17	2020/2/3	2020/2/17		2020/2/3	1.142857143	0.142857143	
18	2020/2/4	2020/2/18		2020/2/4	0.857142857	0.142857143	

図 16　ワークシート:遅れの分析の完成例

図 16 では，B1 に遅れの日数 (X 日) を指定し，感染者数の日付とその X 日後の日付を死者数の日付とします．Excel などの表計算やコンピュータシステムでは，（西暦の）日付に 1 を加えれば 1 日進む（翌日）になるようになっています．そこ

で，分析の開始に日付を 2020/1/22(単純移動平均が求められた最初の日付) とし，A5 に入力します．この A5 の値は分析で必要に応じて変更できるようにします．毎日のサンプルがあるので，下のセルは 1 日後にするので 1 を加え，2021/11/30 まで用意します．

A5:	2020/1/22		
A6:	=A5+1		
複写元:	A6	複写先:	A7:A683

また，遅れの日数 (X の値) を B2 に入力しておきます．遅れを分析するにあたって，死者数を X 日遅れさせた系列を作成します．B5 は, 感染者数の日付の X 日後なので，感染者数の日付に B2 の遅れの日数を加えます．B2 の遅れの日数は複写しても変わらないので絶対参照にします．

B5:	=A5+$B$1		
複写元:	B5	複写先:	B6:B683

図 16 の E 列に A 列の日付の感染者数，F 列に B 列に B 列の日付の死者数を求めます．情報入門 1 で「推移分析」で移動平均値を求めました（図 17）．そのワークシートの E,F 列の値を vlookup 関数を使って取り出し，グラフの横軸として，D 列に感染者数の日付を転記します．日付と感染者数は，推移分析の$D$9:$F$790 にあります．

E5:	=VLOOKUP(A5, 推移分析!$D$9:$F$790,2,FALSE)		
F5:	=VLOOKUP(B5, 推移分析!$D$9:$F$790,3,FALSE)		
D5:	=A5		
複写元:	D5:F5	複写先:	D5:F683

動画:各日付の感染者数と X 日後の死者数の表の作成

	A	B	C	D	E	F
1		原系列	原系列			
2	全国	感染者数	死者数			
3	2020/1/16	1	0			
4	2020/1/17	0	0			
5	2020/1/18	0	0			
6	2020/1/19	0	0			
7	2020/1/20	0	0		単純移動平均	単純移動平均
8	2020/1/21	0	0	全国	感染者数	死者数
9	2020/1/22	0	0	2020/1/22	0.142857143	0
10	2020/1/23	0	0	2020/1/23	0	0
11	2020/1/24	1	0	2020/1/24	0.142857143	0
12	2020/1/25	1	0	2020/1/25	0.285714286	0
13	2020/1/26	1	0	2020/1/26	0.428571429	0

図 17　全国感染者数・死者数推移（情報入門 1）

■グラフ化　情報入門第 5 章,5.5 節と同様に，D4:F683 を第 1 軸を感染者数（折れ線），第 2 軸を X 日遅れの死者数（折れ線）にした組み合わせグラフを作成します．また，縦軸を対数目盛にします．

図 18 は全国感染者数・X 日遅れの死者数推移（図は $X = 14$）の折れ線グラフです．X の値は，B1 に書かれているので，X の値を変更してしてみましょう．また，X の値を変更して，すべての波のピークやボトムがおおよそ一致する X の値を試行錯誤でさがしてみましょう．

動画:各日付の感染者数と X 日後の死者数のグラフ

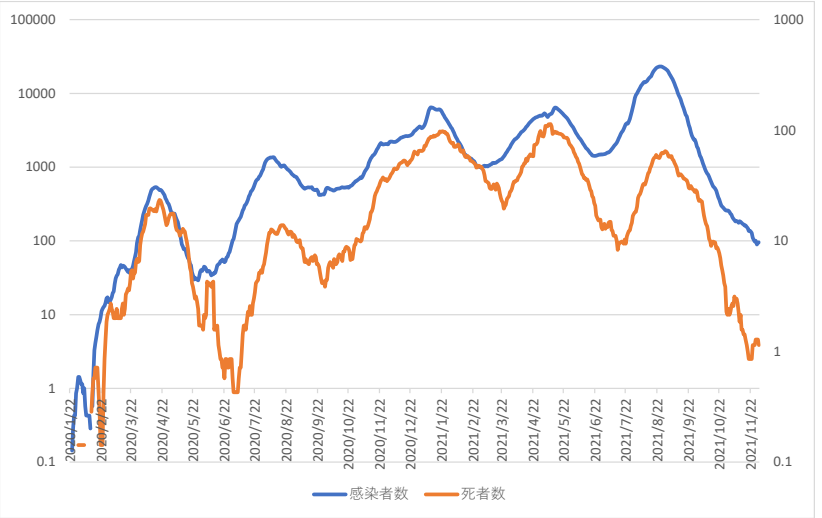


図 18　全国感染者数・X 日遅れの死者数推移（図は $X = 14$）

7.3　散布図と相関係数を用いて遅れの日数を推定

図 18 で，すべての波のピークやボトムがおおよそ一致するような遅れ日数 X を試行錯誤で求めました．これは，感染者数が多い日は X 日遅れの死者数が多い日になるという正の相関関係があると考えられます．この正の相関関係が最も強くなる (相関係数が最大となる) 遅れ日数を遅れ日数の推定値と考えることができるでしょう．

図 19 は，感染者数と X 日遅れの死者数の XY グラフです．

動画:各日付の感染者数と X 日後の死者数散布図

X の値 (B1) 値を変更すると，図 18 の折れ線グラフと散布図が変化します．図 18 のピークやボトムが一致するようにな

ると，図 19 の点は，直線に近づいていくのがわかると思います．ただし，図 19 の点は，2 つの部分 – 急に大きくなる大きな集団（集団 A） と緩やかに上昇する比較的小さな集団 (集団 B) – に別れており，それぞれが直線に近づいていきます．この分析は 7.4 節で行います．

感染者数と X 日遅れの死者数の相関係数を F1 に求めてみましょう．

F1: =CORREL(E5:E683,F5:F683)

動画:感染者数と X 日後の死者数の相関係数の計算

B14 の遅れ日数を変更すると，F1 の相関係数，2 つのグラフの形状が変化します．B14 の値 (非負の整数でおおよそ 30 以下) を変化させて，最大の相関係数の遅れ日数を求めましょう．これを遅れ日数の推定値とします．

7.4　2 つの集団に分けて分析

図 19 を見ると，先ほど指摘したように，2 つの点の集団に分かれているのではないかと書きました．緩やかに上昇する集団（集団 B）の日付をみると，2021 年の 7 月頃から 11 月頃です．これは，図 18 をみると第 5 波の日付であることが分かります．そこで，感染者数の日付で 2021/6/21 までを集団 A(感染者数のボトムで判断してみました)，2021/6/22 以降を集団 B として分析してみましょう．

(1)「遅れの分析」のワークシートをコピーして「集団 A」とします．

(2) A5 に，分析の開始日を入力します．
　ここでは，2020/1/22 のままにします．

(3) 2021/6/22 以降の移動平均値を削除します．
　削除: D522:F683

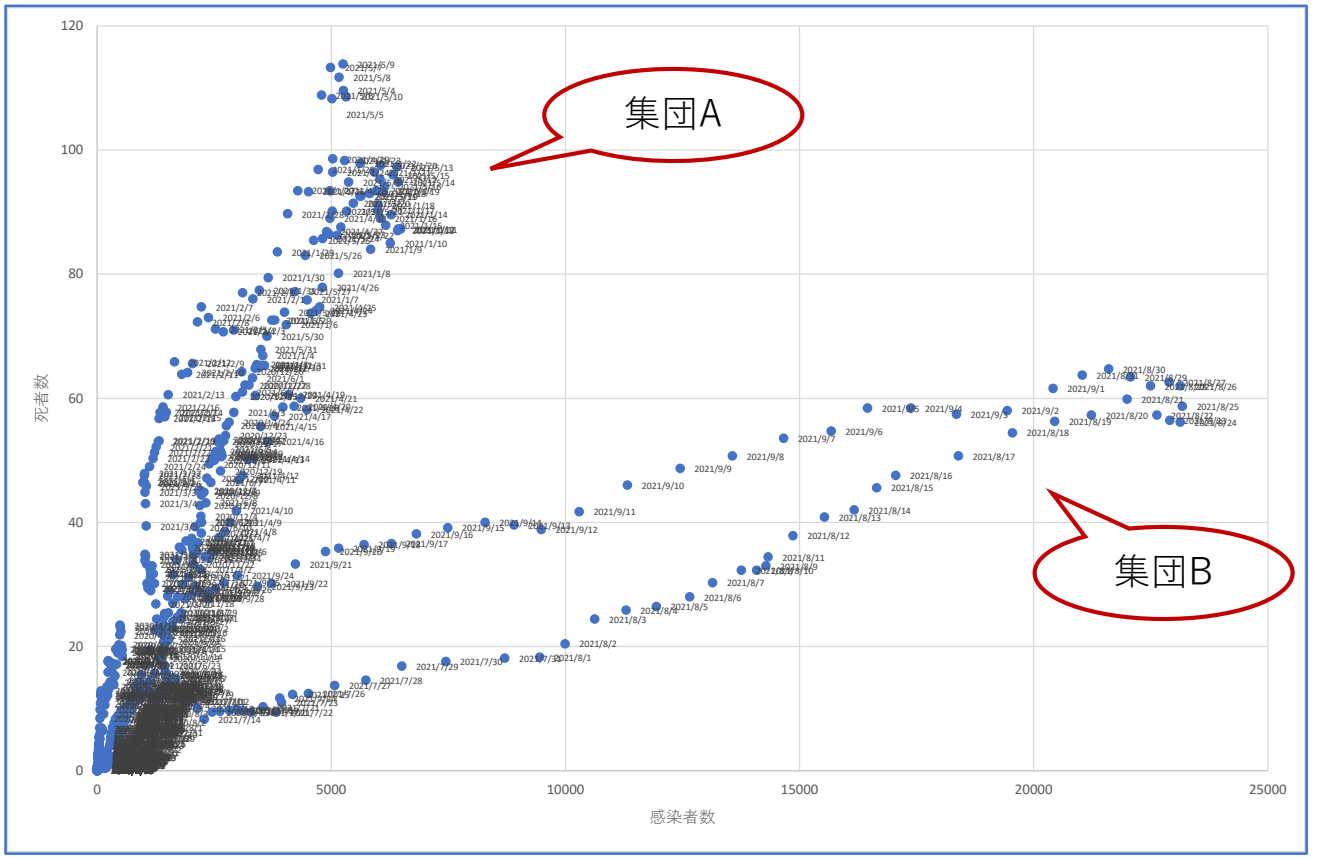


図 19　感染者数と X 日遅れの死者数 XY グラフ（$X = 14$）

2 つのグラフが再描画され，第 4 派までのグラフになります．

(4) 相関係数の計算式を第 3 派までにします．

F1:　=CORREL(E5:E521,F5:F521)

集団 B についても，同様に「遅れの分析」をコピーして，ワークシート「集団 B」を作成し，分析してみましょう．ただし，分析の開始日は 2021/6/22 で，終了日は 2021/11/30 としてみましょう．

8　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）

情報入門 1 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) の平均の支出額について分析をしてみます．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

図 20 は，取得した家計調査ファイルです．図 1 は，図 20 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）までを大 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，情報入門 1 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: A4:G55

動画:家計調査，データの整理

	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 20　取得した家計調査ファイル (ワークシート名:1)

9　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いですし，後ほど，人間が使いやすくなります．

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 21）．

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
- 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．

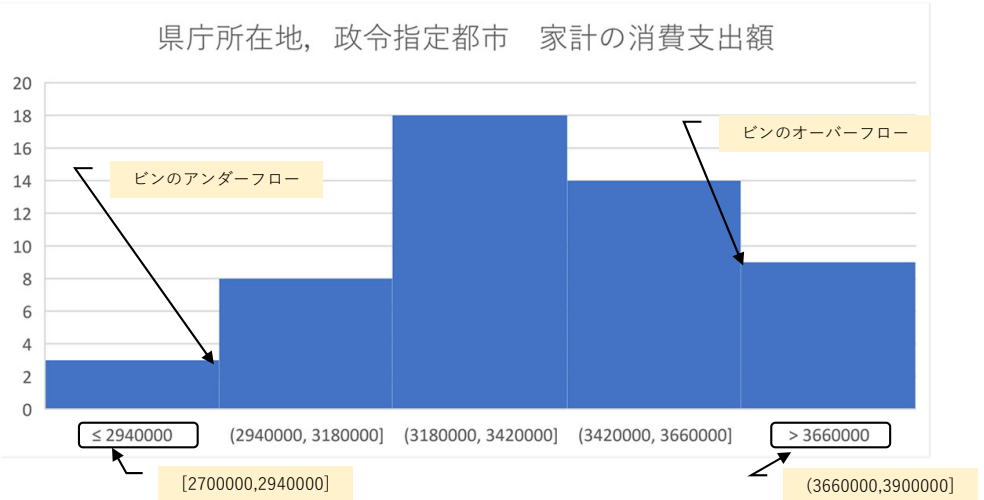


図 21　きりの良い階級のヒストグラム

- 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

	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 22　きりの良い階級を求める

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 22 の C66 の値 (240000)

(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，「≤ 294000」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.