

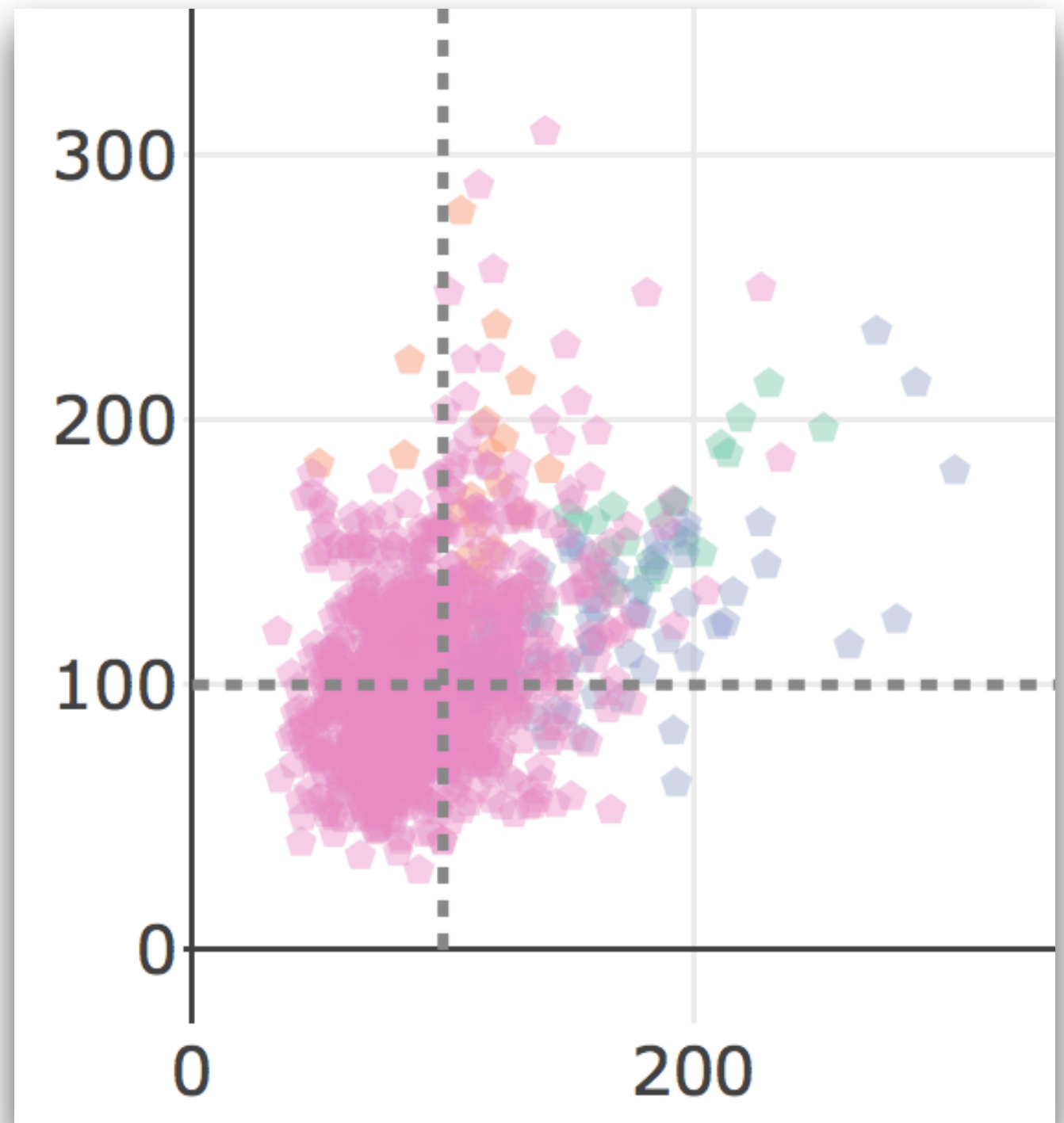
散布図から始めよう

ExcelとRの練習 データ分析の入り口まで

このスライドの最新版とデータは
github.com/nishioWU/JNPC
にあります

散布図とは

- ペアになっているデータを
 - XY平面に図示して
 - 関係を探るのに使う
- ▶ 直線的な関係があるか？
- ▶ 関係は強いかわいいか？
- ▶ トレンドやその変化を見る
- ▶ 外れ値に注目する



なぜ散布図か

- ① データを入手する
- ② 付き合わせたり表記の揺れを直したり加工する
- ③ 散布図を描く
- ④ 相関係数を計算したり、回帰直線を引いたりしてみる

という流れの真ん中に当たる。データの入手・加工の練習にもなるし、分析の糸口にもなる。いろいろ勉強したくなる（のでは）

データのペアの例

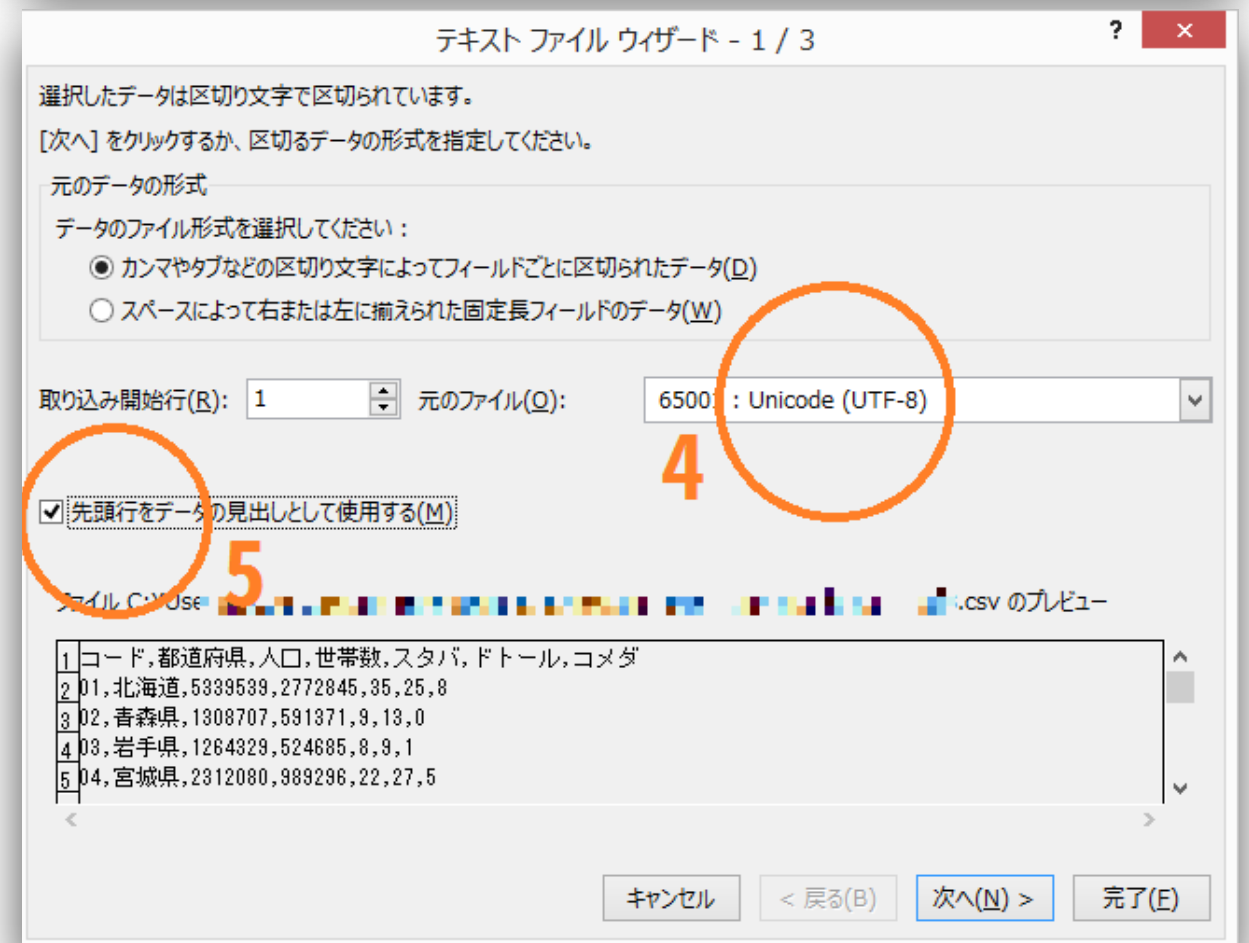
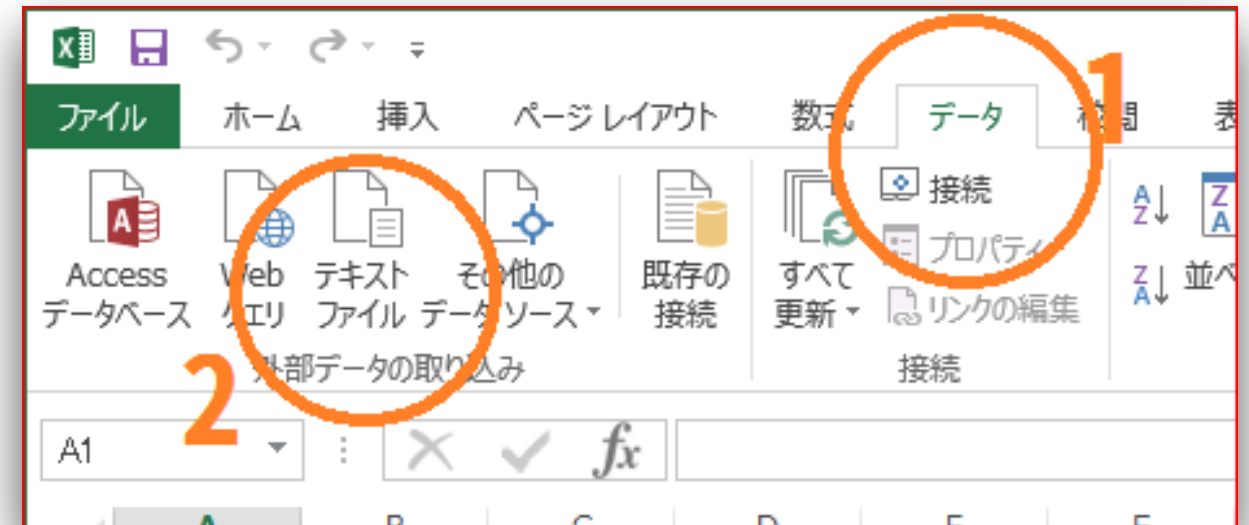
- 身長と、体重
- 親の身長と、子の身長
- 気温と、電力需要
- 気温と、ビール消費量
- 男性の合格率と、女性の合格率
- 自治体の人口と、ある疾患の死亡率
- 各県の人口と、喫茶店の店舗数

Excelに読み込む前に

- 元データはコピーして別名で保存しておく（フォルダ丸ごとが楽）
- 形式「CSV」は、コンマ区切りのテキストのこと
- テキストエディタで中身を確認しておくとい
- 文字コード違いで化ける。UTF-8か、それともS-JIS（CP932）か
- 先頭の0を削るなど、Excelが気を利かせて勝手に変換するのが困りもの。日付も要注意。年がない場合は今年にされてしまう
- なので、ファイルのダブルクリックで開くのはダメ
- さきに空っぽのExcelファイルを作ってからデータを取り込む

CSVデータ読み込み 1

- ① 左上「データ」タブで
- ② 「テキストファイル」を選択
- ③ CSVファイルを選ぶ。今回は、dataフォルダにあるcafe_data.csv。すると「開く」が「インポート」に変わるので、そのボタンをクリック
- ④ 「Unicode (UTF-8)」を選ぶ
- ⑤ 「先頭行をデータの見出しとして使用する(M)」に✓
- ⑥ 「次へ」



CSVデータ読み込み 2

- ① 「区切り文字」を変更。「カンマ」に✓を
- ② プレビュー画面を確認して「次へ」
- ③ 列のデータ形式を指定。「コード」の列を選択してから、上のラジオボタン●で「文字列」にする
- ④ 0で始まるデータに注意。「標準」だと数値に変換され0が消えてしまうので、「文字列」にしておく
- ⑤ 年の入っていない日付も、勝手に今年にされてしまう。もしあれば、やはり「文字列」が安全
- ⑥ 今回は「コード」列以外はデフォルトで、「完了」
- ⑦ データの貼り付け先を適宜指定し、「OK」

CSVデータ読み込み 2 続き

テキスト ファイル ウィザード - 2 / 3

フィールドの区切り文字を指定してください。[データのプレビュー] ボックスには区切り位置が表示されます。

区切り文字

- ☐ タブ(T)
- ☐ セミicolon(M)
- ☒ カンマ(C)
- ☐ スペース(S)
- ☐ その他(O):

連続した区切り文字は 1

文字列の引用符(Q): "

データのプレビュー(P)

コード	都道府県	人口	世帯数	スタバ	ドトール
01	北海道	5339539	2772845	35	25
02	青森県	1308707	591371	9	13
03	岩手県	1264329	524685	8	9
04	宮城県	2312080	989296	22	27

テキスト ファイル ウィザード - 3 / 3

区切ったあとの列のデータ形式を選択してください。

列のデータ形式

- ☐ G/標準(G)
- ☒ 文字列(I)
- ☐ 日付(D): YMD
- ☐ 削除する(I)

[G/標準] を選択すると、数字は数値に、日付は日付形式の値に、その他の値は文字列に変換されます。

詳細(A)...

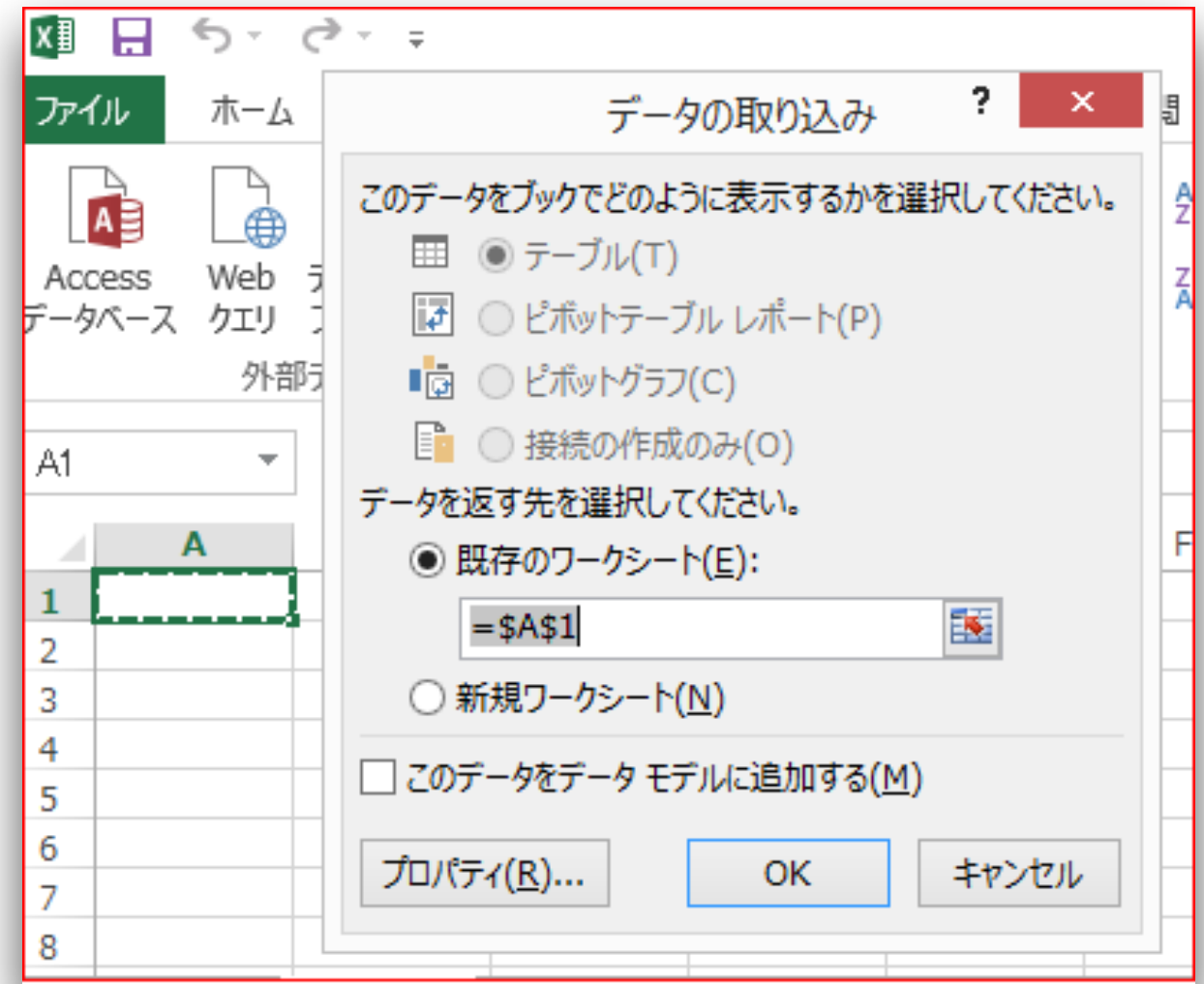
データのプレビュー(P)

文字列	G/標準	G/標準	G/標準	G/標準	G/標準	G/標準
コード	都道府県	人口	世帯数	スタバ	ドトール	コメダ
01	北海道	5339539	2772845	35	25	8
02	青森県	1308707	591371	9	13	0
03	岩手県	1264329	524685	8	9	1
04	宮城県	2312080	989296	22	27	5

キャンセル < 戻る(B) 次へ(N) > 完了(E)

データ読み込み完了

- OKを押せば、シートの左上隅を起点に貼りつく
- 左上隅以外や、新しくシートを増やして貼る場合は、その旨指定を



ここまで済んだ状態：[cafe excel 01.xlsx](#)

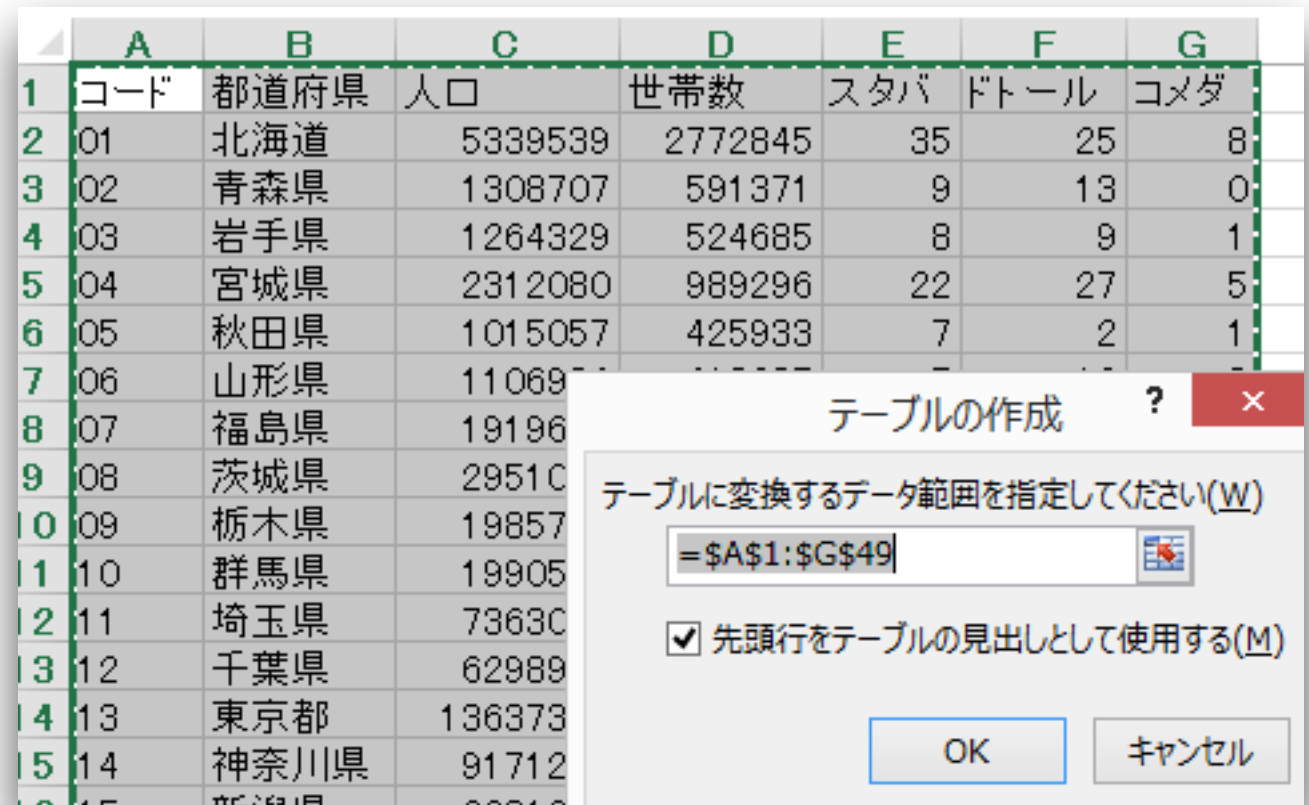
絞り込み可能な表にする

- ① Excelのシートを複写しておく。失敗した場合のリカバリー用。
簡単なやり方は、Ctrl+ドラッグ
- ② データに通し番号を打つ。今回は「コード」で代用して省略
- ③ 絞り込み・ソート可能な表にすると便利。Ctrl+A または Ctrl+*
で範囲指定してCtrl+Tが早い

▶人口順とかコメダが多い順
とか、並べ替え可能になった

▶元に戻すときは、通し番号
の列で「昇順」に

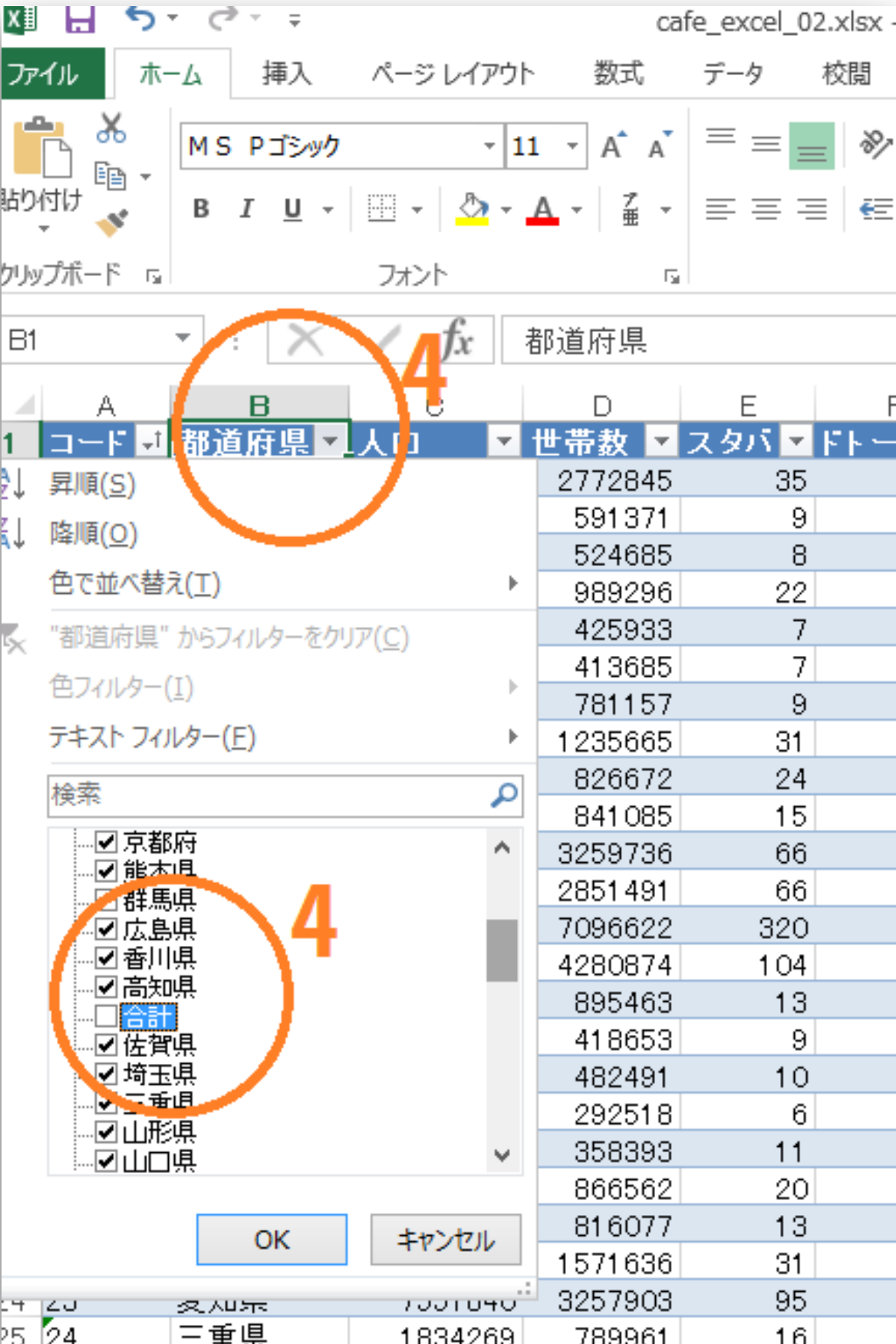
▶フィルターが散布図と連動
するので便利



表にする 2

④ 「都道府県」列の「合計」、または「コード列」の「NULL」を除外しておく、散布図を描くときに困らない。「都道府県」の列でフィルターを使い、「合計」だけ✓を外しておく。これで、全国計が表示されなくなる

表にしておくと、後がラク



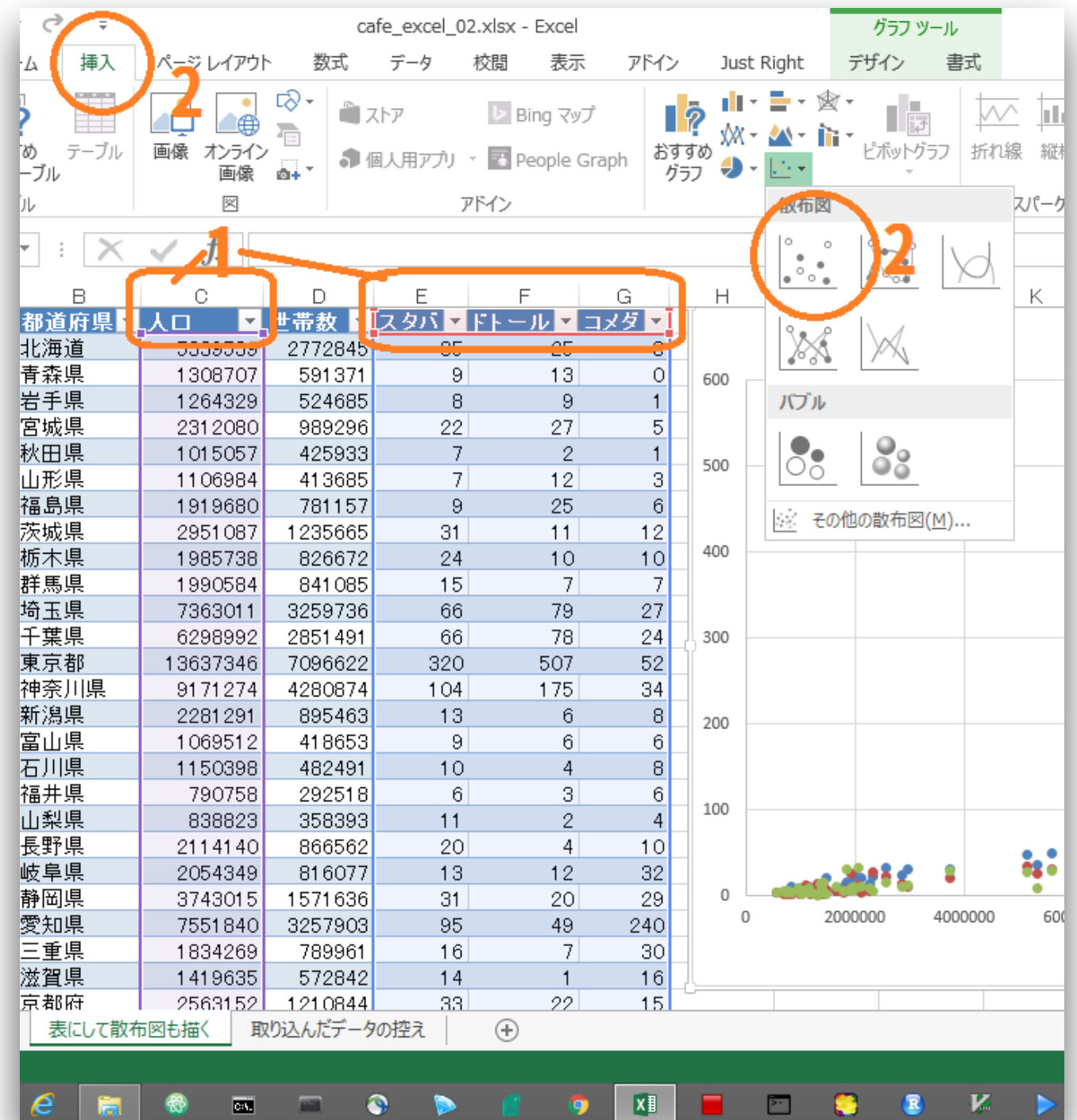
The screenshot shows an Excel spreadsheet with the following data:

コード	都道府県	人口	世帯数	スタバ	ドトー
昇順(S)		2772845	35		
降順(O)		591371	9		
色で並べ替え(I)		524685	8		
"都道府県" からフィルターをクリア(C)		989296	22		
色フィルター(I)		425933	7		
テキスト フィルター(E)		413685	7		
検索		781157	9		
<input checked="" type="checkbox"/> 京都府		1235665	31		
<input checked="" type="checkbox"/> 熊本県		826672	24		
<input checked="" type="checkbox"/> 群馬県		841085	15		
<input checked="" type="checkbox"/> 広島県		3259736	66		
<input checked="" type="checkbox"/> 香川県		2851491	66		
<input checked="" type="checkbox"/> 高知県		7096622	320		
<input type="checkbox"/> 合計		4280874	104		
<input checked="" type="checkbox"/> 佐賀県		895463	13		
<input checked="" type="checkbox"/> 埼玉県		418653	9		
<input checked="" type="checkbox"/> 三重県		482491	10		
<input checked="" type="checkbox"/> 山形県		292518	6		
<input checked="" type="checkbox"/> 山口県		358393	11		
		866562	20		
		816077	13		
		1571636	31		
		3257903	95		
		1834269	789961	16	

散布図を描く 簡単な方法

- ① 表にある列を2つ選ぶ。2列目が離れているなら、1列目を選んだ後、Ctrl+クリックで追加
- ② 「挿入」タブから「グラフ」の「散布図」を選ぶ
- ③ もし3列以上選んだ場合は、左端の列がX座標、残りはY座標になる

では、トライ！ 人口と世帯数で、まず練習してみてもいい



散布図を描く 手動でやるなら

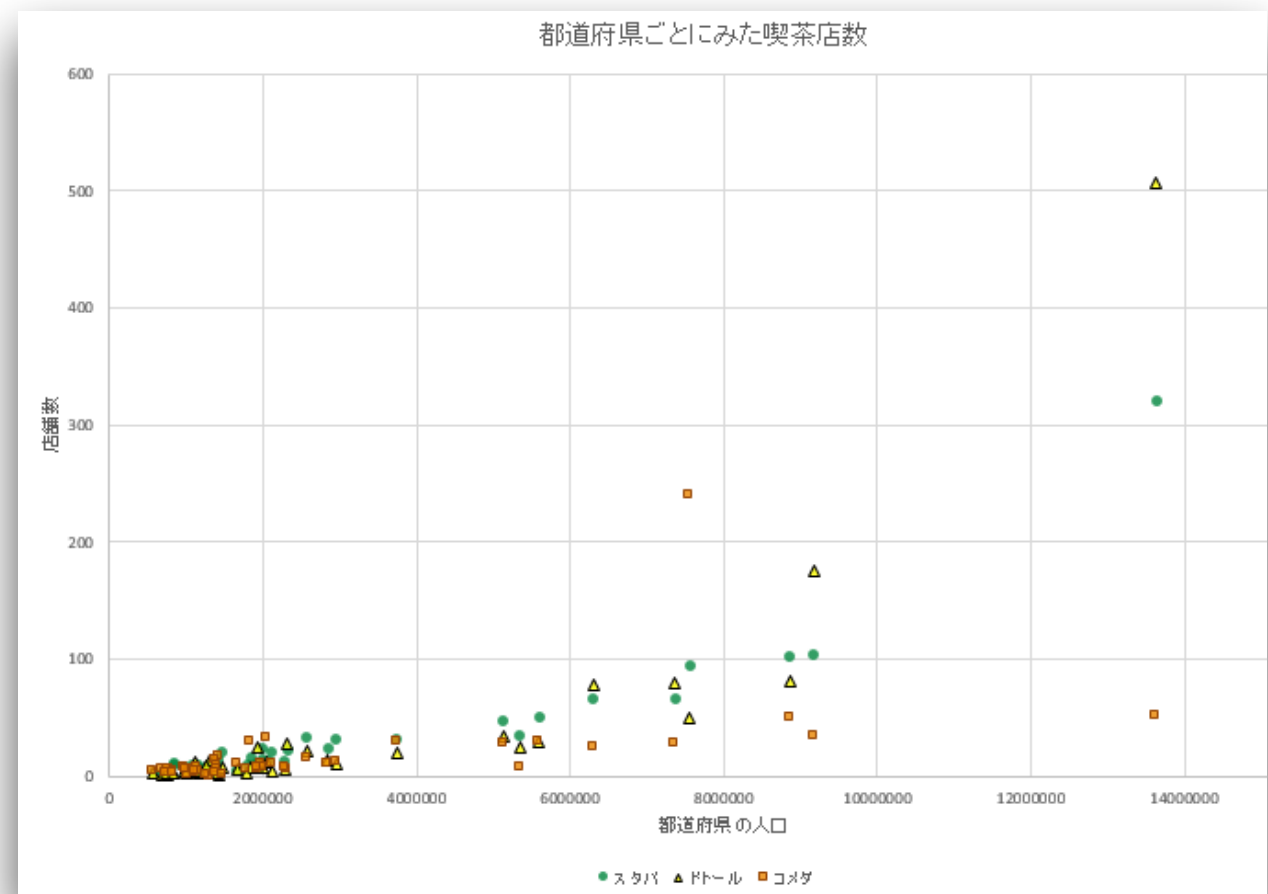
- ① 今あるグラフを手直しするか、列を選ばずに空っぽのままの散布図を挿入。グラフエリア右の「漏斗」のアイコンから、「データの選択」に進み、左側の「凡例項目」の窓で指示する
- ② 「追加」で開く「系列の編集」パネルで、上から順に項目名（じか打ちしても、入っているセルを指定しても可）、X座標のデータ範囲、Y座標のデータ範囲、を指定。XとYは先頭のイコールを残しておく。「OK」で完了
- ③ やり直す場合は「編集」。不要な列があれば「削除」

実は・・・体裁変更が面倒

- 色や形の変更は、点を右クリックして「データ系列の書式設定」「マーカー」と進み、「マーカーのオプション」から
- 都道府県名を添えたければ、グラフエリア右の「+」アイコン → 「データラベルの書式設定」から「セルの値」で都道府県名の列を指定。「Y値」の✓は外す

体裁を直してみた例：

cafe excel 02.xlsx



表にデータ列を追加

グラフの元データの表には、データ列をさらに追加できる。入力済みのテーブルの隣の列に新たに何か入力すると、表の範囲が自動的に広がる仕組み。現有データをもとに計算して追加することも可能。

たとえば、人口の多い県に店が多くても不思議はないので、人口10万人あたりに直したスタバ店舗数を使おう、と考えたら……

- ① 「H2」セルをクリックして「=」だけ打つ
- ② 「E2（スタバ店舗数）」セルをクリックして、「/」を打つ。「C2（人口）」セルをクリックして、「*100000」と打つ
- ③ 列見出しの「H1」を適宜付け直す

これを使って新たに散布図を描けばよい

ここまでを反映：

[cafe excel 03.xlsx](#)

参考になるサイト

- ・「エクセル2016 散布図グラフの作り方」

2016以外のバージョンでも参考になる。マーカー（点）の色の変更、軸の目盛りなど、細かな設定の解説あり

<https://www.tipsfound.com/excel/05036>

Excelのグラフ全般については

- ・「エクセル2016 グラフの作り方」

<https://www.tipsfound.com/excel/05001>

文字化けしたとき IEの小技巧

CSVかTXT形式のデータが文字化けしているときは、エンコードの違いが原因。UTF-8かSHIFT-JISを試してみる。実はInternet Exploreでコードを変換して保存し直す手がある。割と役立つ。

- ① 拡張子が「.csv」の場合は「.txt」に変えて保存。ピリオドまで消さないように注意
- ② IEでファイルを開く。化けていれば、画面を右クリックしてエンコードを直す。たいてい、自動認識してくれる
- ③ 保存したい形式（上記2通りのどちらか）を選び、別名で保存。別名にしないと、元が消えてしまう
- ④ 必要なら拡張子を「.csv」に戻す。戻さなくても、表計算ソフトに読み込むことは可能

Excelの勘どころ

▶行と列

横を行、縦を列とかカラムと呼んで区別している。

▶式は小文字で

Excelは（Rと違って）大文字でも小文字でも命令を聞いてくれる。なので、関数は小文字で入力するとよい。正しく認識されれば大文字に変換される。小文字のまま残ったら、打ち間違いだと分かる。

▶絶対参照

式をコピーすると、気を利かして、計算対象の行や列をずらしてくれる。それが便利だからだが、困る場合もある。そのときは、ずらされては困るものに「\$」マークをつけると、コピー先でもずれない。これが絶対参照。式の入力中に「F4」キーを押すと、行と列の両方またはどちらかに、\$がついたり消えたりして切り替えられる

能率が上がるショートカット

▶ Ctrl + ドラッグ

シート名のタブをつかみながらだと、そのシートのコピーを作成

▶ Shift + ドラッグ

行や列を選択し、その境目をつかみながらだと、並び替え

▶ Ctrl + 1

セルの書式設定。エルではなくて数字の一（テンキーの1はダメ）

▶ Ctrl + Z

直前の変更を元に戻す

▶ Ctrl + A

シート全体を選択

▶ Ctrl + C

コピー

ショートカットその2

▶ Ctrl + V

通常の貼り付け。セル幅以外すべて引き継ぐ。もう一度押すと、貼り付けの形式を選ぶ

▶ Alt + Ctrl + V

形式を選択して貼り付け。Ctrl + Vでは困るときに使う。関数を使って整形をした後、貼り直して「値だけ」にするのに便利（Vを選ぶ）

▶ Alt + ;

絞り込み時に表示されているセルだけをコピー元にする。重宝する

▶ Ctrl + S

ファイルを上書き保存

▶ 「F12」

ファイルを別名で保存

ショートカットその3

- ▶ 「F2」

セルの編集

- ▶ Shift + Ctrl + @

セルの表示を「処理の結果」か「数式そのもの」か切り替える

- ▶ Ctrl + F

検索

- ▶ Ctrl + H

置換

- ▶ Ctrl + カーソルキー

空白セルは飛ばし、その次にデータの入っているセルにジャンプ

- ▶ Ctrl + Home

A1セル（左上）にジャンプ

ショートカットその4 (完)

- ▶ Ctrl + End

データの入っている最終セルにジャンプ

- ▶ Ctrl + ;

きょうの日付を入力。便利

- ▶ Ctrl + :

現在の時刻を入力。便利

- ▶ Ctrl + *

データが入っている範囲を選択。離れ小島は選択されない

- ▶ Ctrl + T

テーブルにする

- ▶ Ctrl + Enter

複数のセルに同じデータを入れる。一括して修正するときに便利

- ▶ Alt + 下矢印

そのカラムに入力済みのデータのリストから選ぶ

計算モデルを当てはめる

Y軸の喫茶店数を、X軸の人口を使った数式で計算・説明できないだろうか？ 準備として、人口を万人単位に直したものを作る（やらなくてもよいのだが、数式の係数の桁数をそこそこ確保するため）。

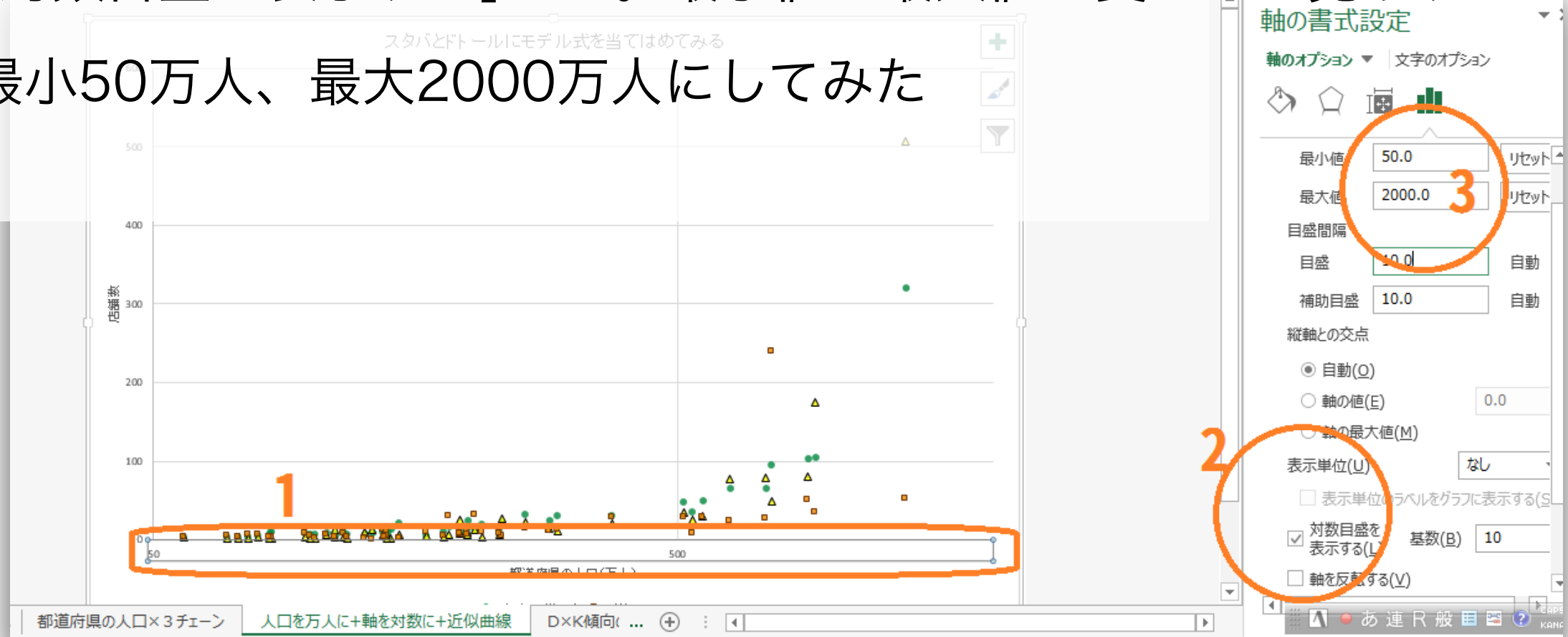
直線的な分布なら、 $y = a + bx$ の一次関数で表せるのでは？

- ① グラフの元表に 1 列追加。都道府県の人口を10000で割った数が入るようにする。15ページの方法の応用
- ② 都道府県人口と 3 チェーンの店舗数をプロットしたシートを複写。Ctrl+ドラッグで。名前を適宜付け替える
- ③ 13ページの方法で各チェーンの系列データを編集。X軸の値だけ、①で作った万人単位に差し替える。あとは触らず

軸を対数に変更する

人口と喫茶店数の関係は、直線的なものではなさそう。対数軸にしてみる。つまり、 $\log y = a + b \log x$ を試してみることにする

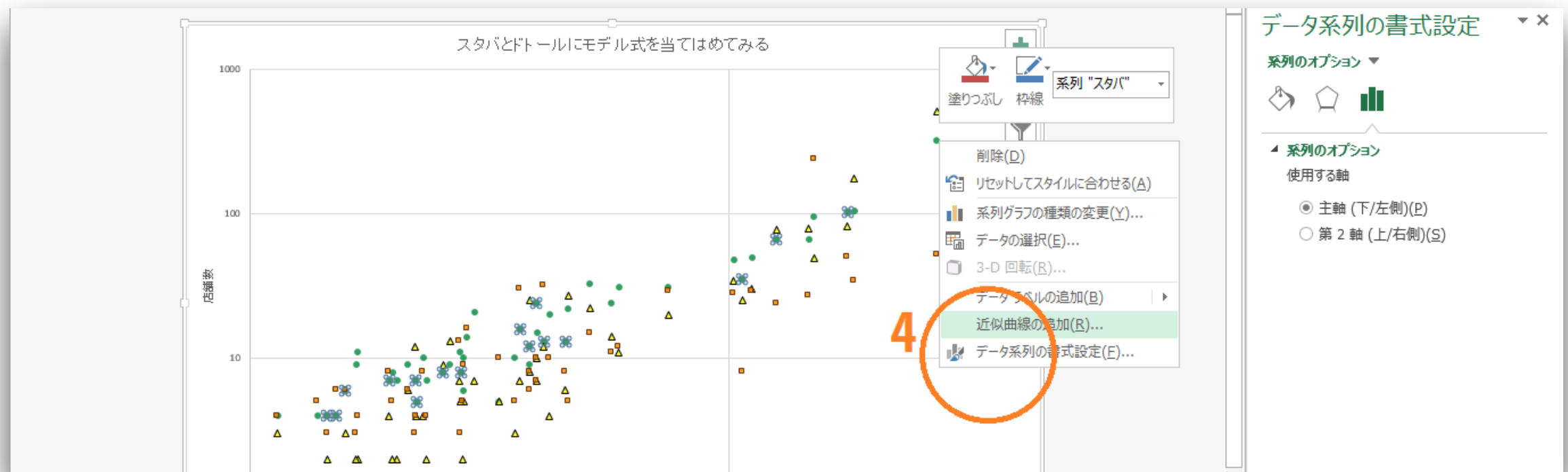
- ① X軸をクリック。軸の書式設定パネルが開く
- ② 「対数目盛を表示する」に✓。最小値・最大値も変えると見やすい
- ③ 最小50万人、最大2000万人にしてみた



近似曲線を引く 1

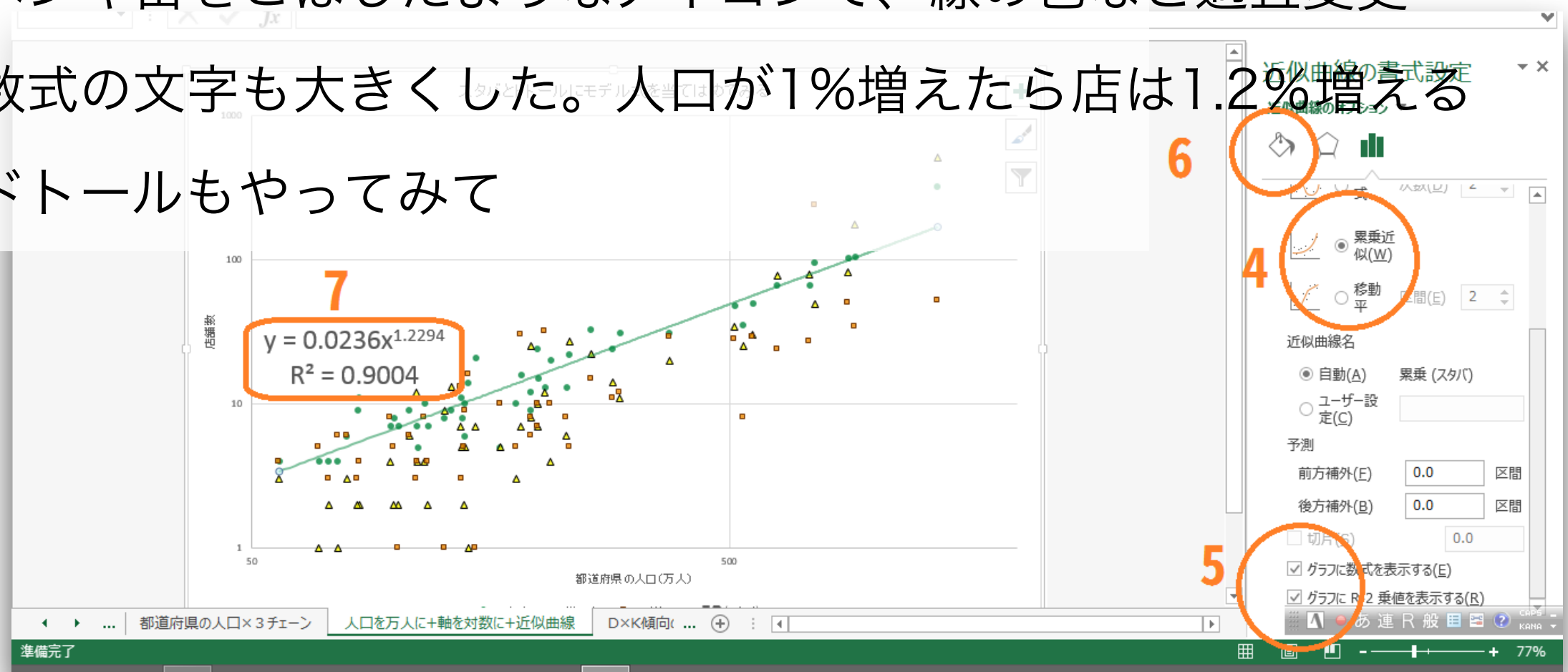
Y軸も対数目盛に変更する。「負の値や0は……」という警告が出るかもしれない。これは、青森にコマダがないため。そのまま続行。だいたい直線的になったので、近似曲線を追加してみる

- ④ スタバのマーカーのどれかを右クリック。「近似曲線の追加」を選ぶ。いきなり線が引かれてしまったら、右クリックして書式変更。線が引かれても、引かれなくても、次ページへ



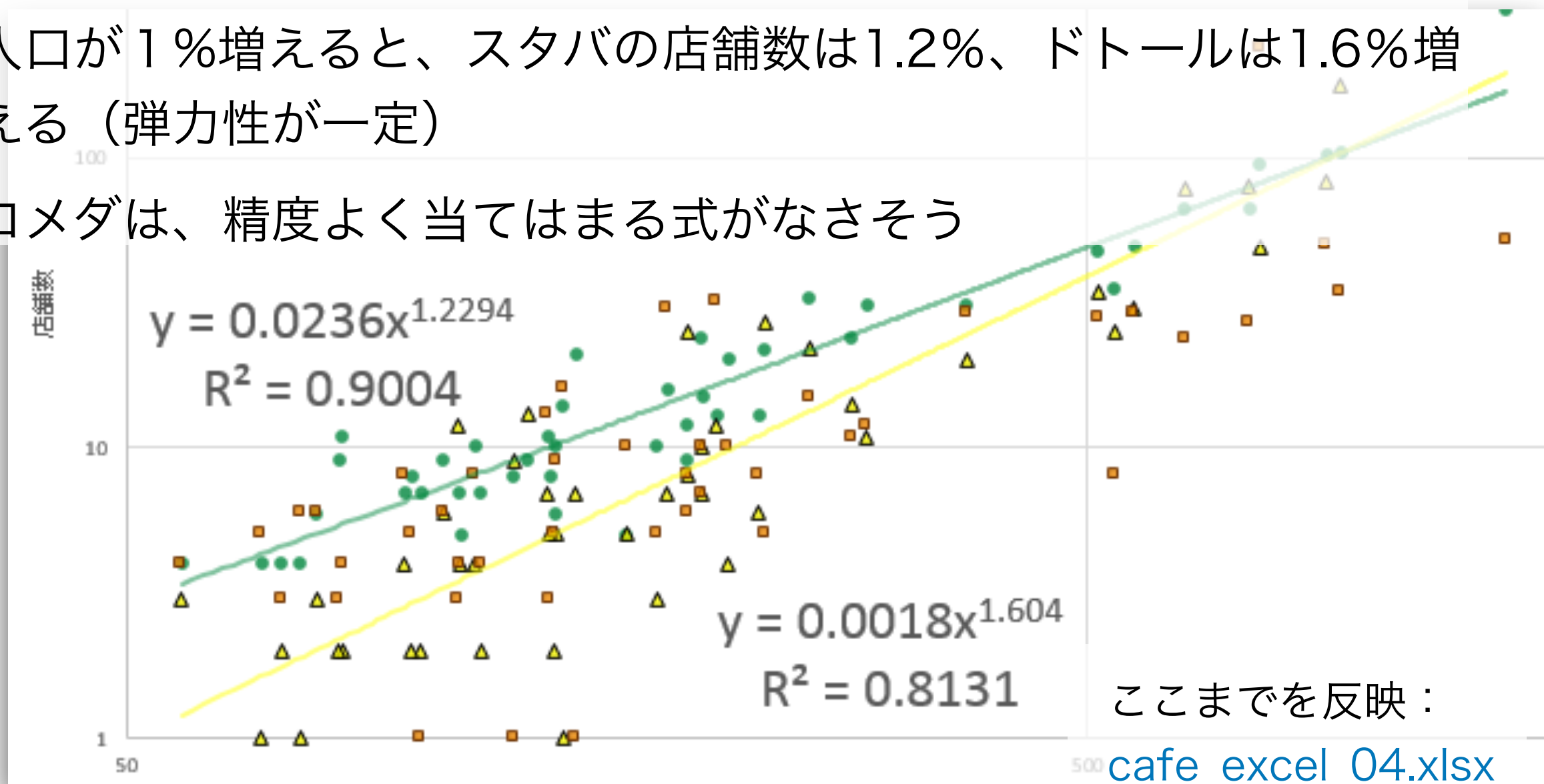
近似曲線を引く 2

- ④ 続き。ボタンで「累乗近似」を選ぶ。対数軸なので直線状になる。
なるべく多くのマーカーの近くを通る線が引かれる
- ⑤ 「グラフに数式を表示」「R2乗値を表示」の両方に✓
- ⑥ ペンキ缶をこぼしたようなアイコンで、線の色など適宜変更
- ⑦ 数式の文字も大きくした。人口が1%増えたら店は1.2%増える
- ⑧ ドトールもやってみて



近似曲線を引く 3 (完)

- ・「R2乗値」とは、Xの式でYの値をどれくらい説明できているかの目安。1に近いほど、よく近似できている
- ・人口が1%増えると、スタバの店舗数は1.2%、ドトールは1.6%増える（弾力性が一定）
- ・コメダは、精度よく当てはまる式がなさそう



相関係数を計算する

散布図を眺めて、X軸とY軸のデータに直線的な関係がありそうなら、相関係数を計算してみる。 $-1 \sim 1$ の値になる。

ExcelではCORREL関数、Rならcor関数を使う。Excelではアドインの分析ツールを有効にして、それを使ってもいい。

- 絶対値が1に近ければ、強い関係
- 0なら相関なし

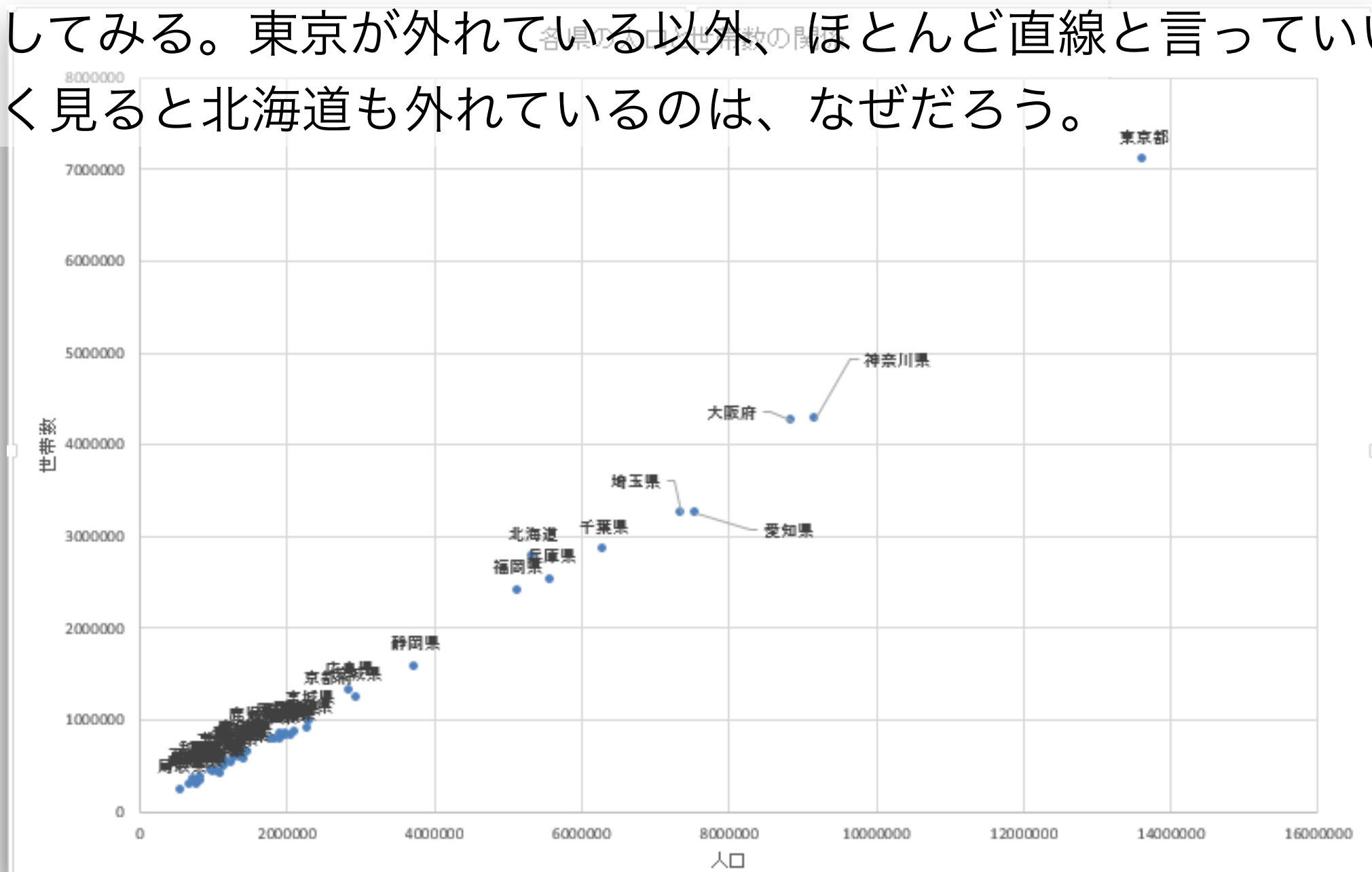
で、関係の強弱の目安になる。散布図を描きもしないで、相関係数を求めるのは、実は危ない。

ただし散布図を眺めてから

直線的な関係ではなさそうなら、相関係数を求めるのはナンセンス。
機械的に計算したら、以下の図はすべて相関係数が同じ。

強い相関の例

喫茶店データにある、都道府県の人口の列と、世帯数の列を散布図にしてみる。東京が外れている以外、ほとんど直線と言っていい。よく見ると北海道も外れているのは、なぜだろう。



CORREL(X, Y)関数

- ① 適当なセルに「=correl(」と入力
- ② X要素の範囲、この場合は「c2:c48」を指定。打ち込んでも、マウスやカーソルで範囲指定してもよい。C列全部を意味する「c:c」にはしないこと。散布図で除外した全国計が含まれてしまう
- ③ コンマで区切り、Y要素の範囲「d2:d48」を指定する
- ④ 「)」を閉じて改行。相関係数は0.99586と計算された

The screenshot shows an Excel spreadsheet with the following data:

コード	都道府県	人口	世帯数	人口(万)
01	北海道	5339539	2772845	533.9539
02	青森県	1308707	591371	130.8707
03	岩手県	1264329	524685	126.4329
04	宮城県	2312080	989296	231.208
05	秋田県	1015057	425933	101.5057
06	山形県	1106984	413685	110.6984
07	福島県	1919680	781157	191.968
08	茨城県	2951087	1235665	295.1087
09	栃木県	1985738	826672	198.5738
10	群馬県	1990584	841085	199.0584
11	埼玉県	7363011	3259736	736.3011

The formula bar shows: `=CORREL(C2:C48,D2:D48)`

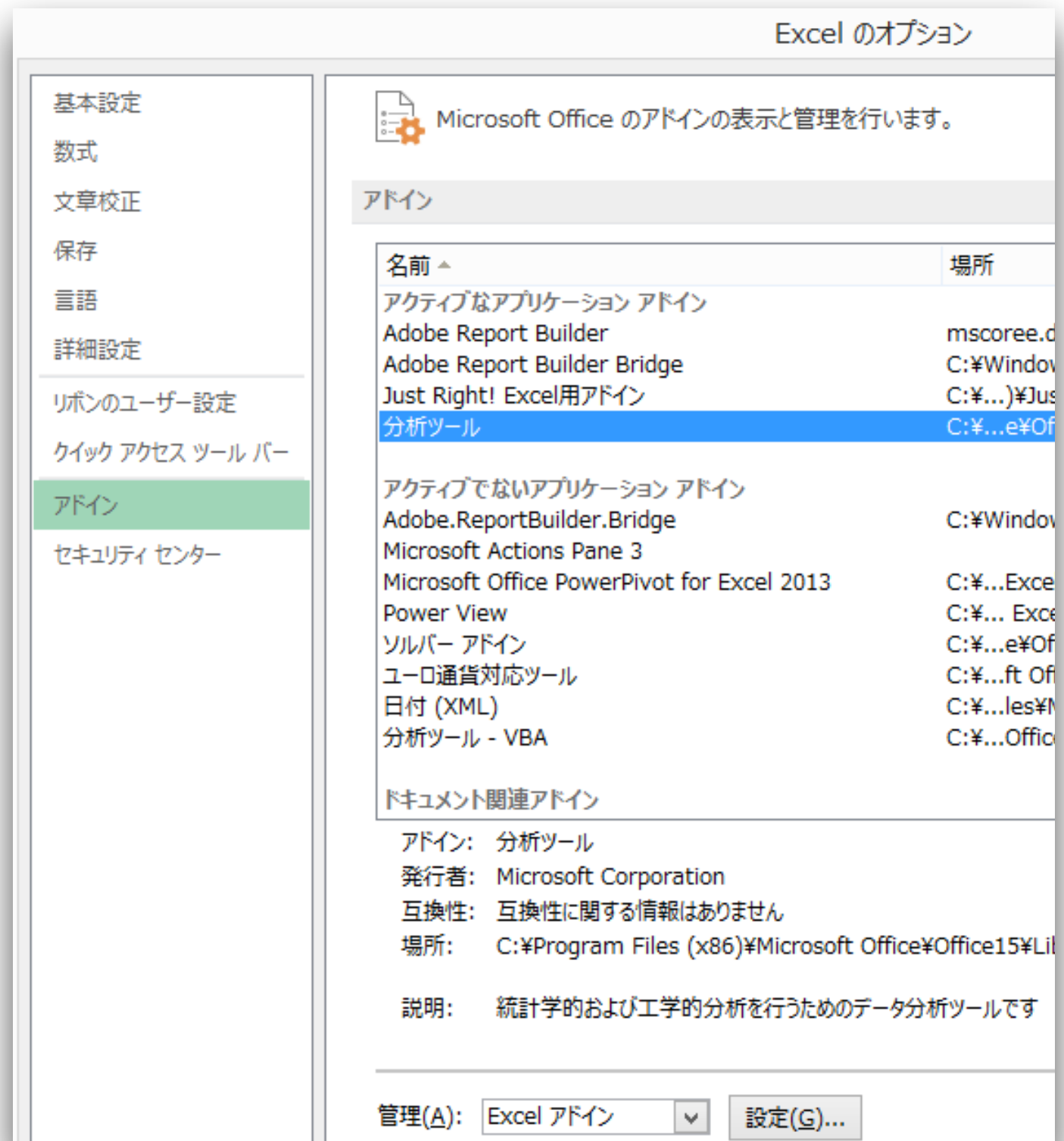
The result of the formula is: **0.99175**

解説: 人口(C列)と世帯数(D列)の相関係数を求める。PEARSON関数と同じように、範囲を指定する。ここでは、C2:C48とD2:D48を指定している。相関係数は、2つのデータの相関の強さを示す値で、-1から1の範囲にある。1に近い値は強い正の相関を示し、-1に近い値は強い負の相関を示す。0に近い値は相関が弱いことを示す。

分析ツールのアドイン

デフォルトではツールが使えないので、使えるようにする（一度だけやればOK）。

左上の「ファイル」タブから「オプション」→「アドイン」と進み、「分析ツール」を有効にするよう設定



分析ツールを使う 1

- ① 「データ」タブから「データ分析」（「分析」の中にある）を選ぶ
- ② 開いたパネルで「相関」を選ぶ

The screenshot shows the Excel interface with the 'Data' tab selected. The 'Data Analysis' button in the 'Analysis' group is circled in orange. Below the ribbon, the formula bar shows $=CORREL(C2:C48,D2:D48)$. The spreadsheet data includes columns for population, households, and other metrics. A 'Data Analysis' dialog box is open, with the 'Correlation' option selected and circled in orange.

県	人口	世帯数	スタバ	ドトール	コメダ	S人口	D人口	K人口	人口(万)
	5339539	2772845	35	25	8	0.655487	0.468205	0.149826	533.9539
	1308707	591371	9	13	0	0.687702	0.993347	0	130.8707
	1264329	524685	8	9	1	0.632747	0.71184	0.079093	126.4329
	2312080	989296	22	27	5	0.951524	1.16778	0.216255	231.208
	1015057	425933	7	2	1	0.689616	0.197033	0.098517	101.5057
	1106984	413685	7	12	3	0.632349	1.084027	0.271007	110.6984
	1919680	781157	9	25	6	0.468828	1.3023	0.312552	191.968
	2951087	1235665	31	11	12	1.05046	0.372744	0.40663	295.1087
	1985738	826672	24	10	10	1.208619	0.503591	0.503591	198.5738
	1990584	841085	15	7	7	0.753548	0.351656	0.351656	199.0584
	7363011	3259736	66	79	27	0.896372	1.072931	0.366698	736.3011
	6298992	2851491	66	78	24	1.047787	1.238293	0.381013	629.8992
	13637346	7096622	320	507	52	2.346498	3.717732	0.381306	1363.7346
	9171274	4280874	104	175	34	1.133975	1.908132	0.370723	917.1274
	2281291	895463	13	6	8	0.569853	0.263009	0.350679	228.1291
	1069512	418653	9	6	6	0.841505	0.561004	0.561004	106.9512
	1150398	482491	10	4	8	0.869264	0.347706	0.695412	115.0398
	790758	292518	6	3	6	0.758766	0.379383	0.758766	79.0758

データ分析

分析ツール(A)
分散分析: 繰り返しのない二元配置
相関
共分散
基本統計量
指数平均
F検定: 2標本を使った分散の検定
フーリエ解析
ヒストグラム
移動平均
指数平滑

分析ツールを使う 2

- ③ 「入力範囲」はXとYをまとめて指定。項目名の行（1行目）も含めるとよい。含めない場合は「列1」「列2」と表示されるので分かりにくい
- ④ 「先頭行をラベルとして使用」に✓
- ⑤ 「出力先」を指定して、「OK」

関数を使ったときと同じ結果になる。対角線に1が入っているのは、自分自身との相関係数を求めているから。

なお、相関係数の2乗が、近似曲線で出てきたR²乗値になる。

分析ツールを使う 3 (完)

The screenshot displays an Excel spreadsheet with a table of data for various prefectures (都道府県). The columns include population (人口), household count (世帯数), and several ratios (S人口比, D人口比, K人口比, 人口(万)). The '相関' (Correlation) dialog box is open, showing the input range '\$C\$1:\$D\$48' and output range '\$N\$12'. The dialog box has orange circles and numbers 3, 4, and 5 highlighting specific options: 3 for the input range, 4 for '先頭行をラベルとして使用(L)' (Use first row as labels), and 5 for the output range. The background spreadsheet shows columns for population (人口), household count (世帯数), and various ratios, with rows for different prefectures (都道府県).

回帰直線を引く

散布図に近似曲線を追加。軸を触らないままで直線的関係があるので「線形近似」にする。これが回帰直線。数式の字は大きくしている。

