健康統計学 第5回

今回は、相関と回帰（テキスト33〜44ページ）について学習します。

テキスト

- 『やさしい保健統計学 改訂第5版』縣 俊彦著 (南江堂)

今回の内容

1. 相関
2. 回帰
3. Excelで相関係数と回帰直線を計算
4. Excelで散布図と回帰直線を作成

相関 (correlation)

2種類のデータの間に何らかの関係がある場合、統計学的な関係性がみられるときに、「**相関がある**」や「**相関関係がある**」といいます。

- データの大小に関して、一方の値が変わるにつれて、もう一方の値も変わる
    - 身長と体重
    - 収縮期血圧と拡張期血圧

データの尺度と相関関係

データを大雑把に、量的データ（比例尺度、間隔尺度）と質的データ（順序尺度、名義尺度）に分けるときに、データの尺度によって、相関関係を表す指標は異なります。次の表を参考にしてください。

2つのデータの尺度	相関関係を表す指標
量的データ×量的データ	ピアソンの積率相関係数
順位データ×順位データ	スピアマンの順位相関係数
量的データ×質的データ	相関比
質的データ×質的データ	クラメールの連関（関連）係数

この授業では、よく利用される、ピアソンの積率相関係数とスピアマンの順位相関係数を扱います。

相関係数 (correlation coefficient)

相関の種類

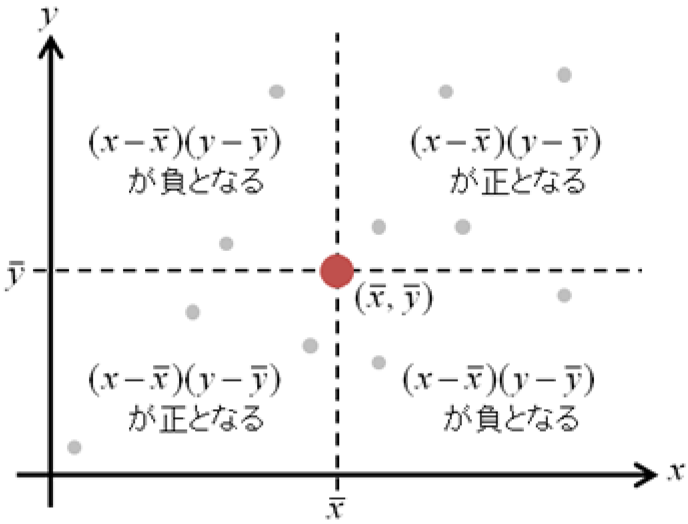
- 線形相関：相関（関係）を示すグラフ（散布図）が1本の直線で近似できる
    - 順相関：相関が正の場合（散布図が右肩あがりの傾向）
    - 逆相関：相関が負の場合（散布図が右肩さがりの傾向）
    - 無相関：相関がない場合（散布図がまばらになっている）
- 非線形相関：相関を示すグラフが指数関数や2次・3次関数のように曲線状になる

偏差積和

- 偏差積和（偏差の積の総和） S_{xy} とは、偏差（各データと平均の差）の積の総和である。

$$S_{xy} = \sum_{i=1}^{n} d_x d_y = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

    - 標本数: n
    - 偏差: d_x , d_y
- 偏差の積とデータの分布は次のようなイメージになる



相関係数（ピアソンの積率相関係数）

- 相関係数（ピアソン(Pearson)の積率相関係数） r は、相関の程度をあらわし、次の値をとる。（一般に相関係数といえばコレ）

$$-1 \leq r \leq +1$$

  - 完全相関：相関係数がちょうど±1の場合
  - 無相関：相関係数が0の場合
- 相関係数 r は、次の式で求められる

$$r = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} = \frac{1}{n}\frac{Sxy}{s_x s_y}$$

  - 標本数: n
  - 標準偏差: s_x , s_y
  - 偏差積和: Sxy
- または、次の式でも求められる（統計量だけから計算できる）

$$r = \frac{1}{s_x s_y}\left(\frac{\sum x_i y_i}{n} - \overline{x}\overline{y}\right) = \frac{1}{s_x s_y}\left(\frac{Txy}{n} - \overline{x}\overline{y}\right)$$

  - 積和（2変数の積の合計）：

$$Txy = \sum_{i=1}^{n} x_i y_i$$

共分散（covariance）

- 共分散 sxy は、偏差積和を標本数で割ったもの。

$$
\begin{aligned}
s_{xy} &= \frac{1}{n} S_{xy} \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n} \sum_{i=1}^{n} d_x \cdot d_y
\end{aligned}
$$

  - 標本数: n
  - 偏差: d_x , d_y

- 共分散 s_{xy} を使うと、相関係数は次のように表せる。

$$
r = \frac{s_{xy}}{s_x s_y}
$$

偏差平方和

- 偏差平方和 S_{xx} または S_{yy} は、偏差の二乗の合計を計算したもの

$$
\begin{aligned}
S_{xx} &= \sum_{i=1}^{n} {d_x}^2 \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}
$$

$$
\begin{aligned}
S_{yy} &= \sum_{i=1}^{n} {d_y}^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2
\end{aligned}
$$

- x と y についての偏差平方和 S_{xx} と S_{yy} を使うと、相関係数は次のように表せる。

$$
\begin{aligned}
r &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \\
&= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}
\end{aligned}
$$

相関関係と因果関係

相関関係から因果関係を確定するときの注意点

何でもよいから2組のデータの関係性を調べればよいわけではありません。次のような5つの因果関係が認められる場合に、相関関係を調べることが有効になります。

1. 関連の時間性
   - 原因は結果の前にあるか
2. 関連の密接性
   - 原因が結果に密接に関連するか

3. 関連の特異性
    - 原因が結果にどの程度かかわっているか
4. 関連の普遍性
    - 対象や時期、方法などが異なっていても、類似した結果が得られるか
5. 関連の合理性
    - 従来の理論や経験から考えて矛盾がないか

疑似相関（見かけの相関）

直接の相関はないが、**何かある要因が2つの事象と相関している**ために、2つの事象に相関がみられるケースがあります。

このような場合を「**疑似相関**」といいます。つまり、相関関係があるからといって、それが必ずしも因果関係であるとは限らない場合です。

- 「ビアホールでの生ビールの売り上げ数」と「アイスクリーム店のお客の数」
    - 2つの事象には「気温」「天候」などが相関している
- 「進行性の疾患をもつ患者の疾患についての知識」と「その疾患の進行度」
    - 2つの事象には「疾患の内容」「治療期間」などが相関している

相関の程度

相関係数の値から、相関の程度を次のように記述できます。

-1.0≦相関係数*r*<-0.7	強い負の相関がある
-0.7≦相関係数*r*<-0.4	かなりな負の相関がある
-0.4≦相関係数*r*<-0.2	やや負の相関がある
-0.2≦相関係数*r*≦0.2	ほとんど相関がない
0.2<相関係数*r*≦0.4	やや正の相関がある
0.4<相関係数*r*≦0.7	かなりな正の相関がある
0.7<相関係数*r*≦1	強い正の相関がある

なお、標本数が少ない場合は、母相関係数の推定や検定（後日説明）が必要となります。

順位相関係数（rank correlation coefficient）

相関がない場合や順位に意味がある・順位だけしかわからない場合には、順位データ（データを小さいほうから並べた順位）をもとに、相関を求める方法が有効になります。

- 英語のテストの順位と数学のテストの順位の相関
- 2つの銘柄の株価の相関（経済分野）
- 薬と奇形児発生の相関（医学分野）

また、順位尺度のデータだけでなく、比例・間隔尺度のデータについても何らかの順位を求めることで適用できます。

- スピアマン（Spearman）の順位相関係数 r_s は、相関係数と同様、次の値をとる。

$$-1 \leq r_s \leq +1$$

- 同一順位の場合は、次のように扱う（平均順位）
    - 2位が2つある場合：2位と3位の中間 (2+3)/2=2.5位を順位とする
    - 4位が3つある場合：4位と5位と6位の中間 (4+5+6)/3=5位を順位とする
- 順位相関係数は、次のようにして求められる。

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n^3 - n}$$

- 標本数: n
- i 番目の順位差: d_i

データ1の順位	データ2の順位	順位差 d	順位差の二乗 d^2
x_1	y_1	$d_1 = x_1 - y_1$	${d_1}^2$
x_2	y_2	$d_2 = x_2 - y_2$	${d_2}^2$
x_3	y_3	$d_3 = x_3 - y_3$	${d_3}^2$
...	...	...	...
x_n	y_n	$d_n = x_n - y_n$	${d_n}^2$
計		0	$\sum_{i=1}^{n} {d_n}^2$

回帰 (regression)

データをもとに、ある変数（**従属変数**または**目的変数**）を別の変数（**独立変数**または**説明変数**）で予測する式を作るための統計的手法を、「**回帰分析**」（regression analysis）といいます。

とくに、独立変数が1つだけの場合を**単回帰分析**といいます。複数の独立変数で1つの従属変数を予測する場合は重回帰分析といいます。

回帰直線 (regression line)

回帰直線

- 散布図の各点 (x_i, y_i) が近くに分布するような直線を回帰直線という。

$$y = ax + b$$

  - 回帰係数 (回帰直線の傾き) : a
  - 回帰直線のy切片 (*x*=0 のときのyの値) : b
  - 独立変数 (説明変数 : 予測に使う変数) : x
  - 従属変数 (目的変数、基準変数 : 予測したい変数) : y
- なお、回帰式は必ず $(\bar{x}, \bar{y})$ を通る

最小二乗法（least squares method）

- 観測値 (または実測値) y_i と 推定値 (または予測値) $\hat{y}_i$ との差 (残差 ϵ) の2乗の和が最小になるような a と b を求める。
  - 次の値が最小となるような、a と b を求める。

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  - 残差の二乗 ϵ^2 を足したものを、残差平方和 S_ϵ という

$$S_\epsilon = \sum \epsilon^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

回帰式の計算

- x を独立変数 (横軸) 、y を従属変数 (縦軸) としたときの回帰式 (y への x からの回帰式) は次のようになる。

$$y = r \cdot \frac{s_y}{s_x} x + \left(\bar{y} - \frac{s_{xy}}{{s_x}^2} \cdot \bar{x} \right)$$

  - なお、回帰式を $y = ax + b$ とすると、 a と b は次のようになる。

$$a = r \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{{s_x}^2}$$
$$b = \bar{y} - \frac{s_{xy}}{{s_x}^2} \cdot \bar{x} = \bar{y} - \bar{x} a$$

  - 相関係数: r
  - 2変数の標準偏差: s_x , s_y
  - 2変数の共分散 (偏差積和の平均)

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

- 回帰式を変形すると、次のようになる。

$$\begin{aligned} y - \overline{y} &= \frac{s_{xy}}{{s_x}^2}(x - \overline{x}) \\ &= \frac{S_{xy}}{S_{xx}}(x - \overline{x}) \end{aligned}$$

  - 2変数の平均値: $\overline{x}$, $\overline{y}$
  - 2変数の偏差積和: S_{xy}

$$S_{xy} = \sum_{i=1}^{n} d_x \cdot d_y = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

  - 偏差平方和: S_{xx}

$$S_{xx} = \sum_{i=1}^{n} {d_x}^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

標準誤差（standard error）

- 予測値と実測値のずれ（予測値 $\widehat{y}$ との誤差；残差 ϵ ）について考える

$$\widehat{y} = a x_i + b + \epsilon$$

- 残差の標準偏差を、標準誤差 s_ϵ という

$$\begin{aligned} s_\epsilon &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y})^2}{n-2}} \\ &= \sqrt{\frac{S_\epsilon}{n-2}} \end{aligned}$$

  - $n-2$ で割っているのは、2つの係数を推定したことによる自由度（後日説明）の修正のため

決定係数（coefficient of determination）

- 相関係数の二乗を決定係数（または寄与率） R^2 という。

$$\begin{aligned} R^2 &= \left(\frac{1}{n}\frac{S_{xy}}{s_x s_y}\right)^2 \\ &= \frac{{S_{xy}}^2}{n^2 \cdot {s_x}^2 {s_y}^2} \\ &= \frac{{S_{xy}}^2}{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2} \\ &= \frac{{S_{xy}}^2}{S_{xx} S_{yy}} \end{aligned}$$

- 偏差平方和と残差平方和を使うと、次のように書くこともできる

$$
\begin{aligned}
R^2 &= \frac{S_{\hat{y}\hat{y}}}{S_{yy}} \\
&= \frac{\sum(\hat{y}_i - \overline{\hat{y}})^2}{\sum(y_i - \overline{y})^2} \\
&= 1 - \frac{S_e}{S_{yy}}
\end{aligned}
$$

- 決定係数は、0から1の値をとる。

$$0 \leq R^2 \leq 1$$

- 推定（回帰式）の精度を表す指標である
  - 「従属変数 y の分散の何%を予測値 $\hat{y}$ の分散が説明しているか」を示す
    - 別の言い方をすると、「説明変数が従属変数の何%にあたる部分に影響を与えているか（寄与しているか）」を示す
  - だいたい、0.5以上であれば精度が高いといえる

回帰の概念

- 予測値 $\hat{y}$ と従属変数の平均 $\overline{y}$ との差は、一般に独立変数 x とその平均 $\overline{x}$ との差より小さくなる
  - →予測値と平均の差 $(\hat{y} - \overline{y})$ との差は、独立変数と平均の差 $(x - \overline{x})$ との差より小さくなる
  - →予測値は独立変数に比べて平均に近づく
- 統計学的な現象で、「回帰効果」や「平均への回帰」ともいう
  - （例）1回目の試験の結果が偏っていた（とくに良い、悪いなど）人について、2回目の試験結果を調べると、その平均値は1回目の結果よりも1回目の全体の平均値に近くなる（時間的には逆で考えてもよい）
  - （例）父親と子どもの身長を比較して、とくに身長の高い父親でも、とくに身長の低い父親からでも、子どもたちの身長は父親たちの身長より平均に近くなる
  - （例）とくに身長の高い人たちの父親の身長は、子どもたちの身長よりも平均に近い（全体の身長の分布は、父親たちも子どもたちも同じ）

Excelで相関と回帰を計算

相関を計算

相関係数

- 2つの配列データの相関係数は、**CORREL**関数を利用します。

CORREL (相関係数の値を返す)

- 書式 : CORREL(配列1, 配列2, ...)
- 引数 : 配列1 ... :データが入力されたセルの範囲
- 引数 : 配列2 ... :もう一方のデータが入力されたセルの範囲
- 例:データがA1～A10セルとB1～B10までのセルの数値から、相関関数を計算する

```
=CORREL(A1:A10, B1:B10)
```

- ピアソンの積率相関係数は、**PEARSON**関数を利用します。

PEARSON (ピアソンの積率相関係数 r の値を返す)

- 書式 : PEARSON(配列1, 配列2)
- 引数 : 配列1 ... :独立変数に対応するセルの範囲
- 引数 : 配列2 ... :従属変数に対応するセルの範囲
- 例:独立変数がA1～A10セル、従属変数がB1～B10までのセルの数値から、積率相関関数を計算する

```
=PEARSON(A1:A10, B1:B10)
```

- なお、Excel2004以降は、CORREL関数の結果とPEARSON関数の結果は同じになります。

共分散

- 共分散 (2種類のデータ間での標準偏差の積の平均値) は、**COVAR**関数または**COVARIANCE.P**関数を利用します。

COVAR (共分散の値を返す)

- 書式 : COVAR(配列1, 配列2)
- 引数 : 配列1 ... :データが入力されたセルの範囲
- 引数 : 配列2 ... :もう一方のデータが入力されたセルの範囲
- 例:データがA1～A10セルとB1～B10までのセルの数値から、共分散を計算する

```
=COVAR(A1:A10, B1:B10)
```

COVARIANCE.P　(共分散の値を返す)

- 書式 : COVARIANCE.P(配列1, 配列2)
- 引数 : 配列1 ... :データが入力されたセルの範囲
- 引数 : 配列2 ... :もう一方のデータが入力されたセルの範囲
- 例:データがA1～A10セルとB1～B10までのセルの数値から、共分散を計算する

```
=COVAR(A1:A10, B1:B10)
```

偏差平方和

- 偏差平方和（標本の平均値に対する各データの偏差の平方和）は、**DEVSQ**関数を利用します。

DEVSQ（偏差平方和の値を返す）

- 書式：DEVSQ(数値1, 数値2, ...)
- 引数：数値1, 数値2 ...：データが入力されたセルの範囲
- 例：データがA1～A10セルのセルの数値から、偏差平方和を計算する

```
=DEVSQ(A1:A10)
```

回帰を計算

回帰直線の傾き

- 既知の y と既知の x のデータから回帰直線の傾きには、**SLOPE**関数を利用します。

SLOPE（回帰直線の傾きを返す）

- 書式：SLOPE(配列1, 配列2)
- 引数：配列1 ...：既知の y（従属変数）に対応するセルの範囲
- 引数：配列2 ...：既知の x（独立変数）に対応するセルの範囲
- 例：既知の y（従属変数）がA1～A10セル、既知の x（独立変数）がB1～B10までのセルの数値から、回帰直線の傾きを計算する

```
=SLOPE(A1:A10, B1:B10)
```

回帰直線のy切片

- 既知の y と既知の x のデータから（線形）回帰直線のy切片には、**INTERCEPT**関数を利用します。

INTERCEPT（回帰直線の切片を返す）

- 書式：INTERCEPT(配列1, 配列2)
- 引数：配列1 ...：既知の y（従属変数）に対応するセルの範囲
- 引数：配列2 ...：既知の x（独立変数）に対応するセルの範囲
- 例：既知の y（従属変数）がA1～A10セル、既知の x（独立変数）がB1～B10までのセルの数値から、回帰直線のy切片を計算する

```
=INTERCEPT(A1:A10, B1:B10)
```

決定係数

- 既知の y と既知の x のデータからR^2（決定係数）を求めるには、**RSQ**関数を利用します。

RSQ（r2の値を返す）

- 書式：RSQ(配列1, 配列2)
- 引数：配列1 ...：既知の y（従属変数）に対応するセルの範囲
- 引数：配列2 ...：既知の x（独立変数）に対応するセルの範囲
- 例：既知の y（従属変数）がA1～A10セル、既知の x（独立変数）がB1～B10までのセルの数値から、決定係数 R2を計算する

```
=RSQ(A1:A10, B1:B10)
```

Excelで散布図と回帰直線を作成

2組のデータの関係性を視覚的に把握するには、「**散布図**」を用います。

また、表計算ソフトのExcelでは、作成した散布図を利用して、「**回帰直線**」（単回帰直線）を描いたり、回帰式を表示することができます。

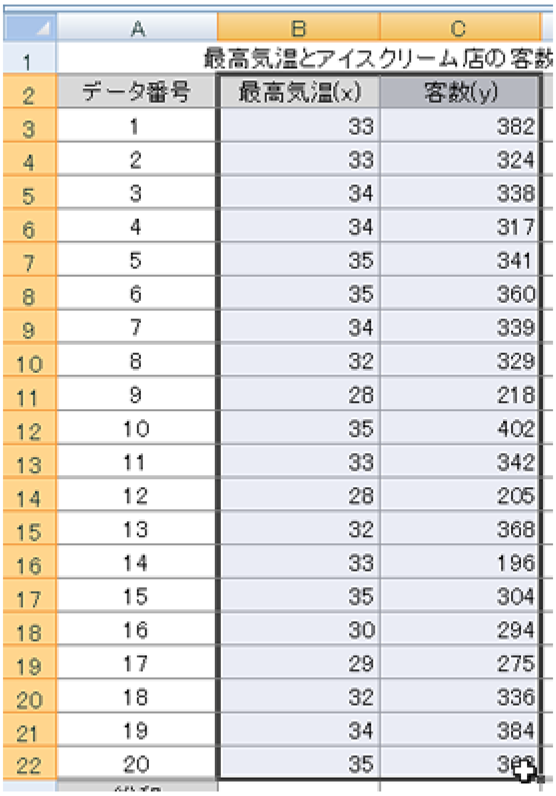
Excelで散布図の作成

散布図の作成

eラーニングの画面からダウンロードできるExcelのファイルを利用してみます。「【練習】相関と回帰」というシートで作成します。

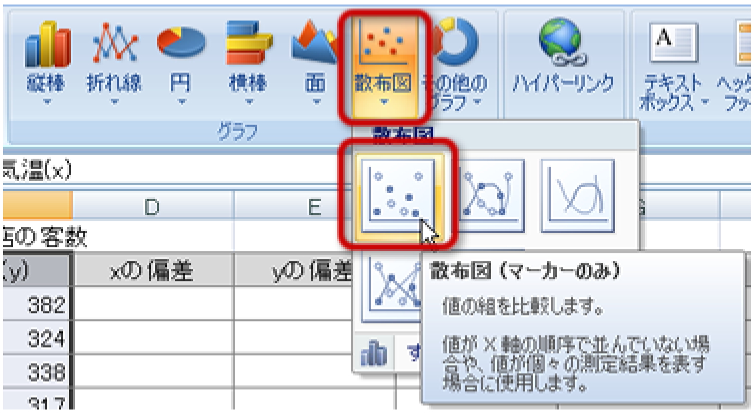
次のようにして、散布図を作成してみましょう。

1. マウスをドラッグして、B2～C22セルを範囲選択する



	A	B	C
1	最高気温とアイスクリーム店の客数		
2	データ番号	最高気温(x)	客数(y)
3	1	33	382
4	2	33	324
5	3	34	338
6	4	34	317
7	5	35	341
8	6	35	360
9	7	34	339
10	8	32	329
11	9	28	218
12	10	35	402
13	11	33	342
14	12	28	205
15	13	32	368
16	14	33	196
17	15	35	304
18	16	30	294
19	17	29	275
20	18	32	336
21	19	34	384
22	20	35	[illegible]

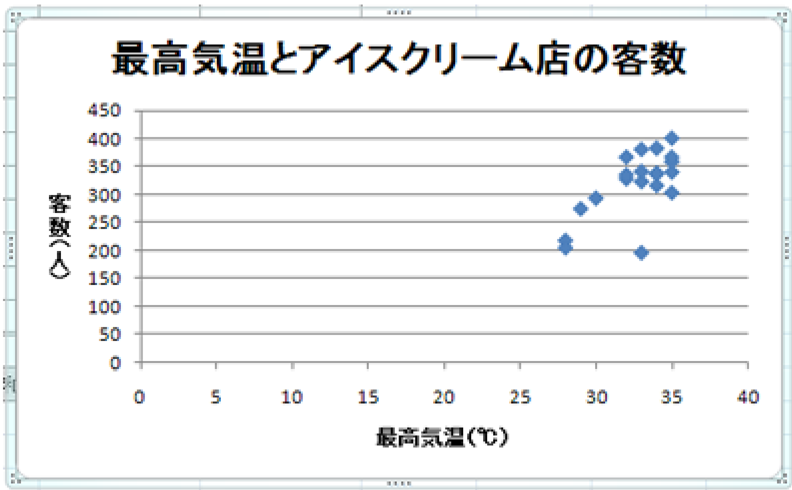
2. 「挿入」タブの「グラフ」グループにある「散布図」ボタンをクリックする
3. メニューから「散布図（マーカーのみ）」を選択する



4. **グラフが作成される**

作成できたら、グラフに次の設定をしてください。

- **凡例は表示を「なし」にする**
- **グラフのタイトルを「最高気温とアイスクリーム店の客数」設定する**
- **縦軸の軸ラベルを「客数(人)」と設定する**
- **横軸の軸ラベルを「最高気温(℃)」と設定する**

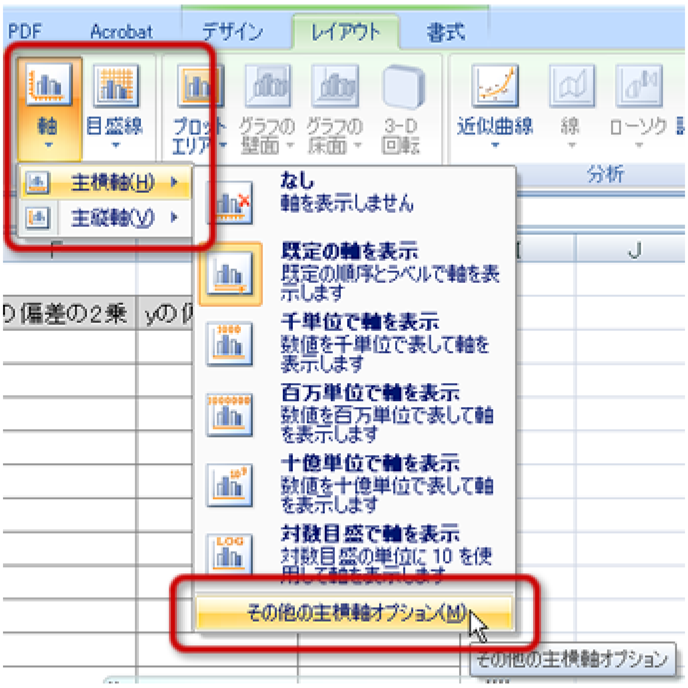


軸の設定

横軸(最高気温)が「0度」から表示されているため、データの分布がグラフの右端に偏ってしまっています。

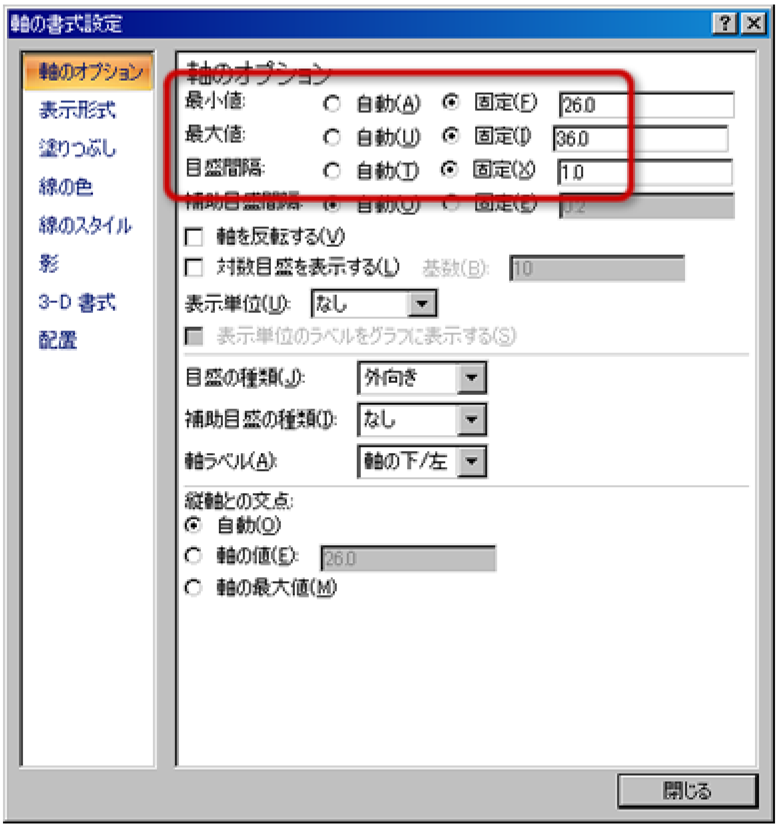
そこで、横軸の範囲を「**26度~36度**」の範囲に変更してみましょう。

1. **メニューの「グラフツール」の「レイアウト」にある「軸」→「主横軸」→「主横軸オプション」を選択する**

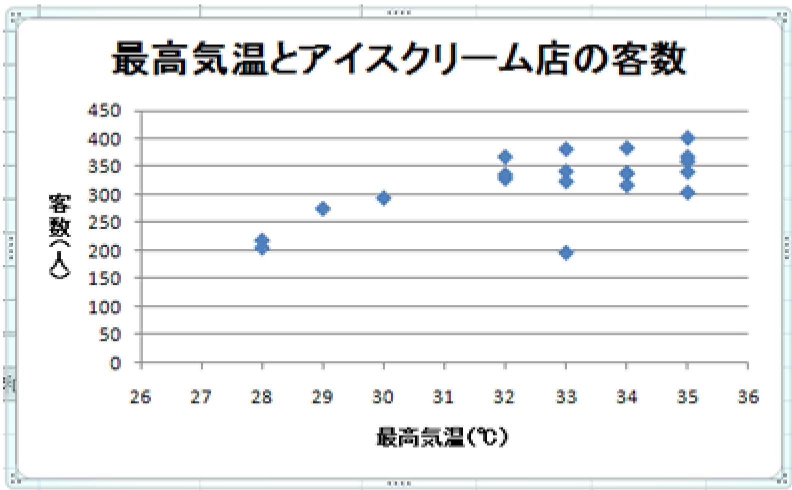


2. 「軸の書式設定」が表示されるので、次のように設定する

- 「最小値」を「固定」に選択して、入力欄に「26.0」と入力
- 「最大値」を「固定」に選択して、入力欄に「36.0」と入力
- 「目盛間隔」を「固定」に選択して、入力欄に「1.0」と入力



3. 「閉じる」をクリックすると、横軸の表示が変更される

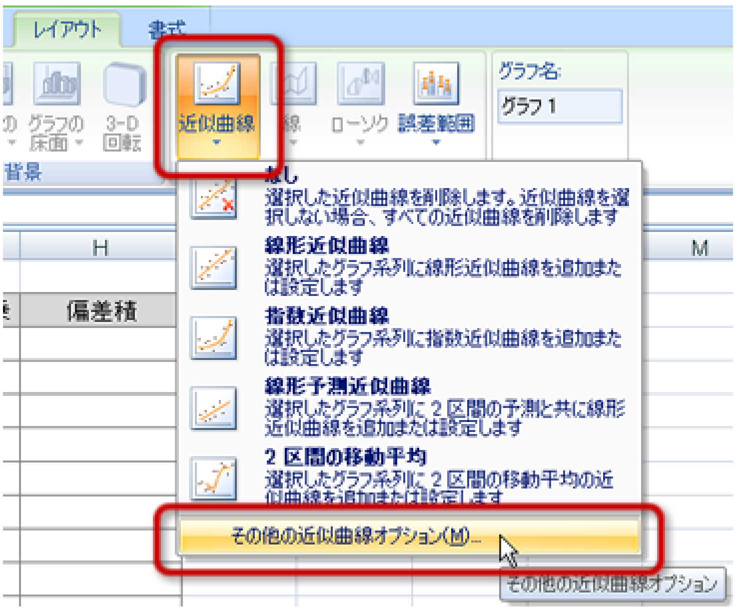


あとは、必要に応じて、縦軸の範囲も設定したり、デザインの変更をすると良いでしょう。

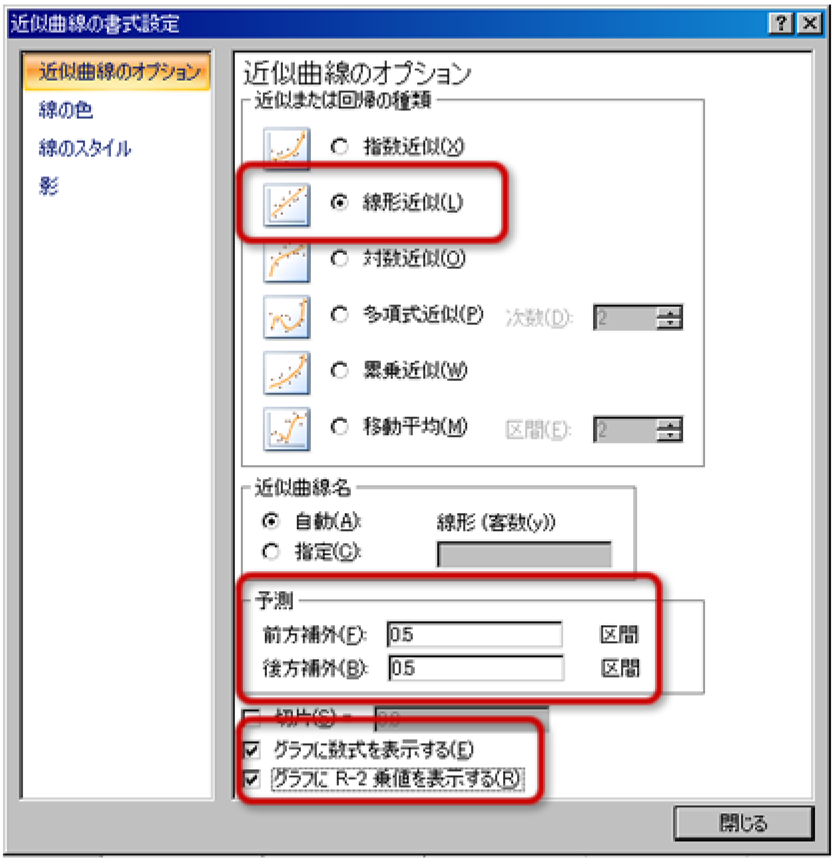
Excelで回帰直線の作成

Excelでは、「近似曲線」という機能で、回帰直線を描くことができます。

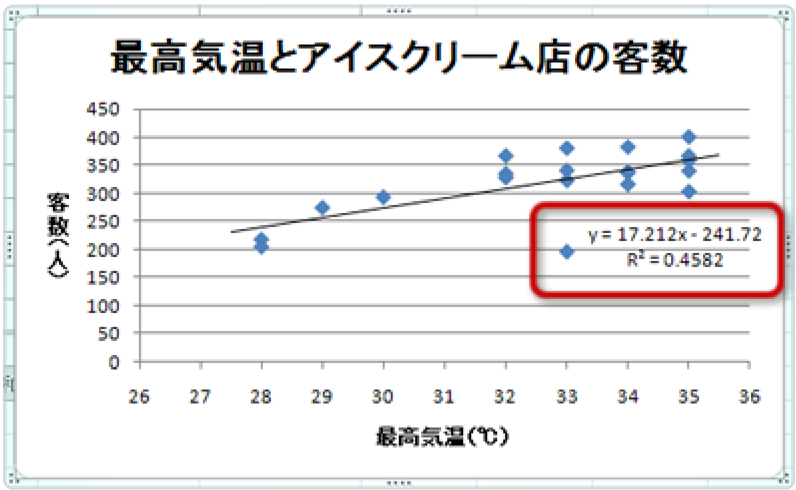
1. メニューの「グラフツール」の「レイアウト」にある「近似曲線」→「その他の近似曲線オプション」を選択する



2. 「近似曲線の書式設定」が表示されるので、次のように設定する
   - 「近似または回帰の種類」を「線形近似」に選択する
   - 「予測」の「前方補外」・「後方補外」の入力欄に、それぞれ「0.5」と入力
   - 「グラフに数式を表示する」と「グラフにR-2乗値を表示する」のチェック欄をクリックしてチェックする



3. 「閉じる」をクリックすると、散布図のグラフに回帰直線（と回帰式）が表示される



これで、回帰曲線と同時に、回帰式（傾き、y切片）、決定係数(R^2値）が表示されます。