2変量データの記述・回帰分析
（教科書第4，5，10章）

北九州市立大学経済学部

齋藤 朗宏

レポート

- 統計的手法の用いられている心理学系の論文を読み，その研究内容と用いられた統計的手法について700 字以上で要約せよ．その上で，用いられている統計的手法について，どういう目的で，どうその手法が用いられているのかなど，300 字以上で論評せよ．尚，論文の掲載されている代表的な学会誌には，以下のようなものがある．（締め切りは8/2，試験の際に同時に回収する）．
  - 心理学研究
  - 教育心理学研究
  - 社会心理学研究

試験

- 8/2に実施する.

- 試験は，検定など実際に計算を行って，その結果について述べる形式.
- PCルーム（この教室）で実施，計算する上でExcelの使用を認める.
- 教科書その他，すべて持ち込み可能だが，一つ一つ調べて回答している時間はないので注意.
- また，試験時間中のインターネットの参照は禁止する.
- 試験時間は60分.

今日の内容

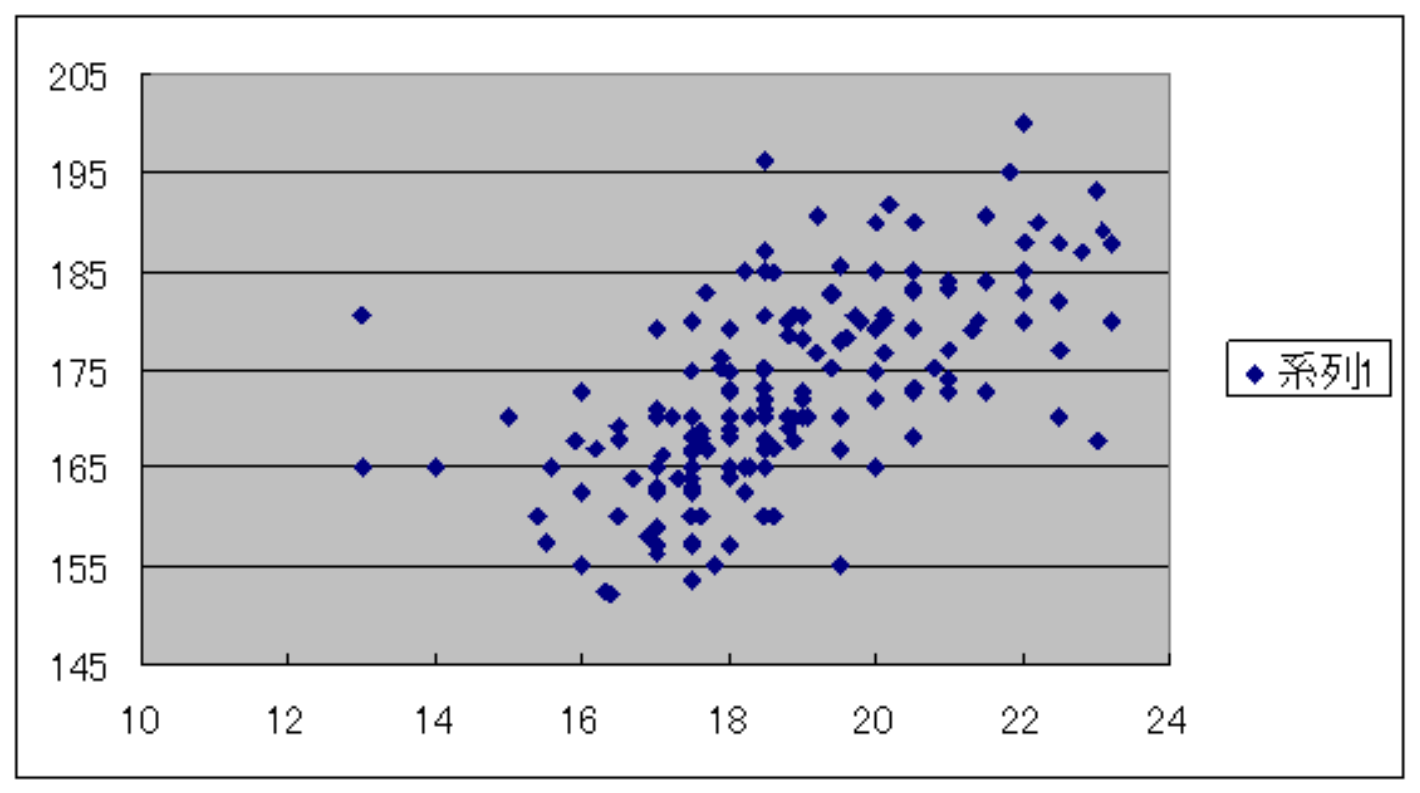
- 散布図（教科書4.3）
- 共分散と相関（教科書5.7）
- 回帰分析（教科書10）

１変量から２変量へ

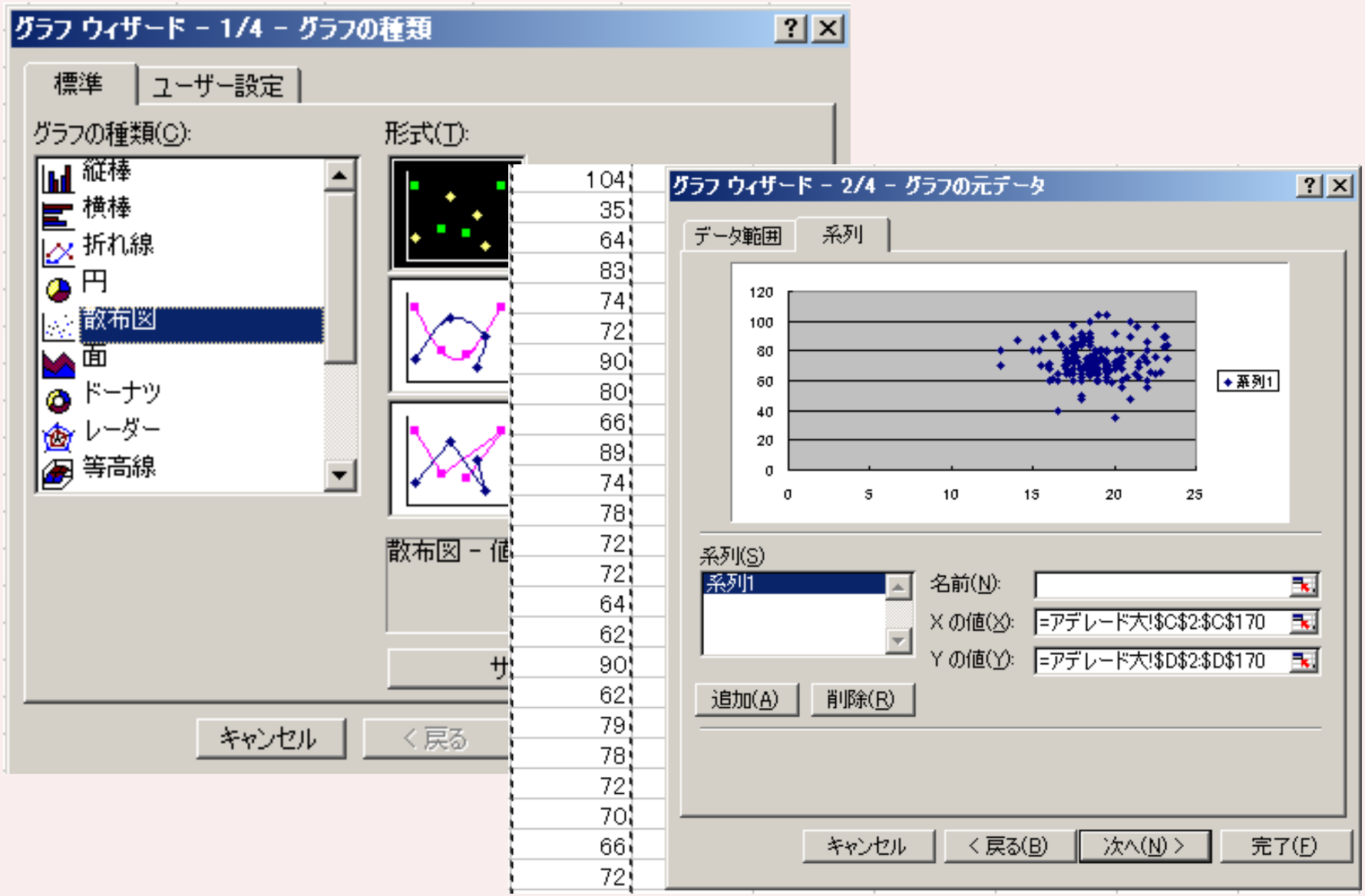
- １変数のデータを要約する場合，平均，標準偏差等を使用すればよかった．
- ２変数になった場合でも，平均や標準偏差等を使用する点はかわらない．
- ただし，それに加えて「変数Aと変数Bの関係性」が興味の対象となる．
- e.g.
  - 身長が高い人は体重も重いことが多い（一方が増えると他方も増える）．
  - 年齢が上がると，テレビの視聴時間が短くなることが多い（一方が増えると他方は減る）．

散布図

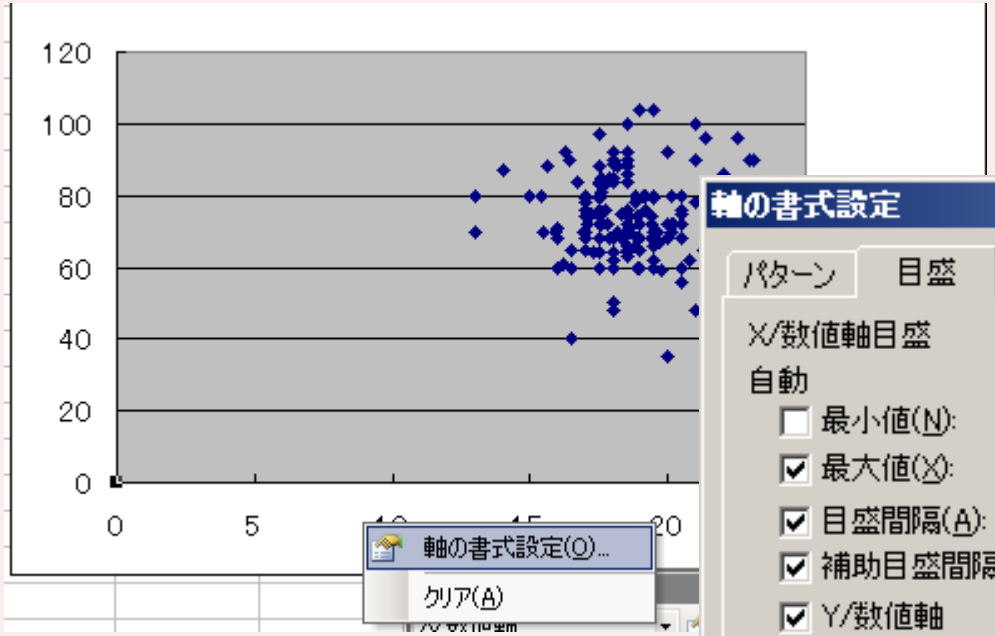
- 主に量的尺度間の関係を見たい場合に用いられる.



Excelを用いた散布図の作成（１）



Excelを用いた散布図の作成（２）



軸の書式設定

パターン | 目盛 | フォント | 表示形式 | 配置

X/数値軸目盛

自動

- [ ] 最小値(N): 10
- [x] 最大値(X): 25
- [x] 目盛間隔(A): 5
- [x] 補助目盛間隔(I): 1
- [x] Y/数値軸との交点(C): 0

表示単位(U): なし

- [x] 表示単位のラベルをグラフに表示する(D)

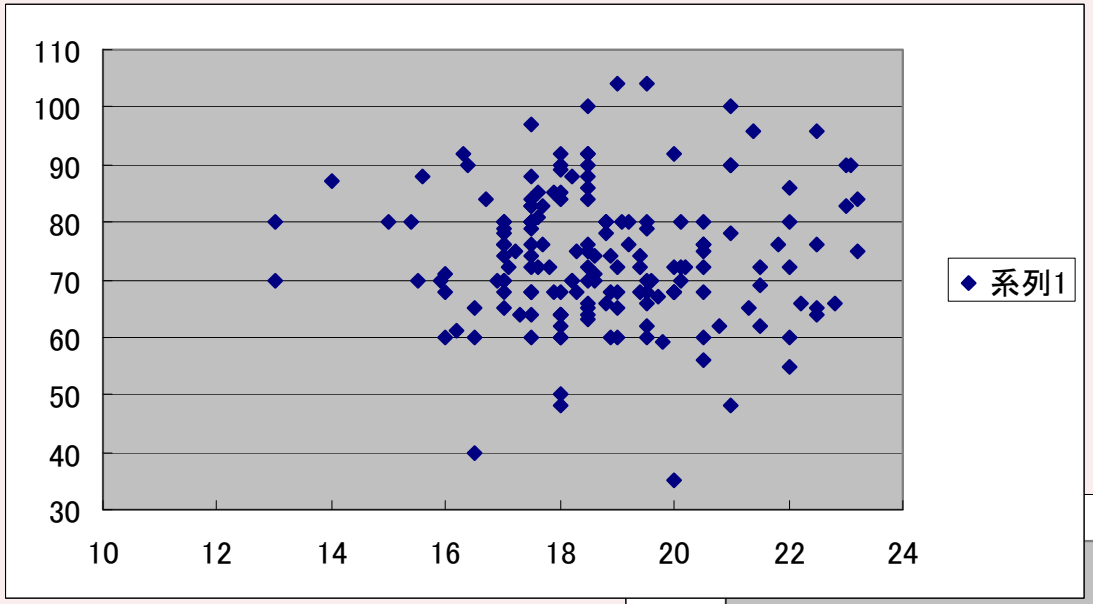
- [ ] 対数目盛を表示する(L)
- [ ] 軸を反転する(R)
- [ ] 最大値で Y/数値軸と交差する(M)

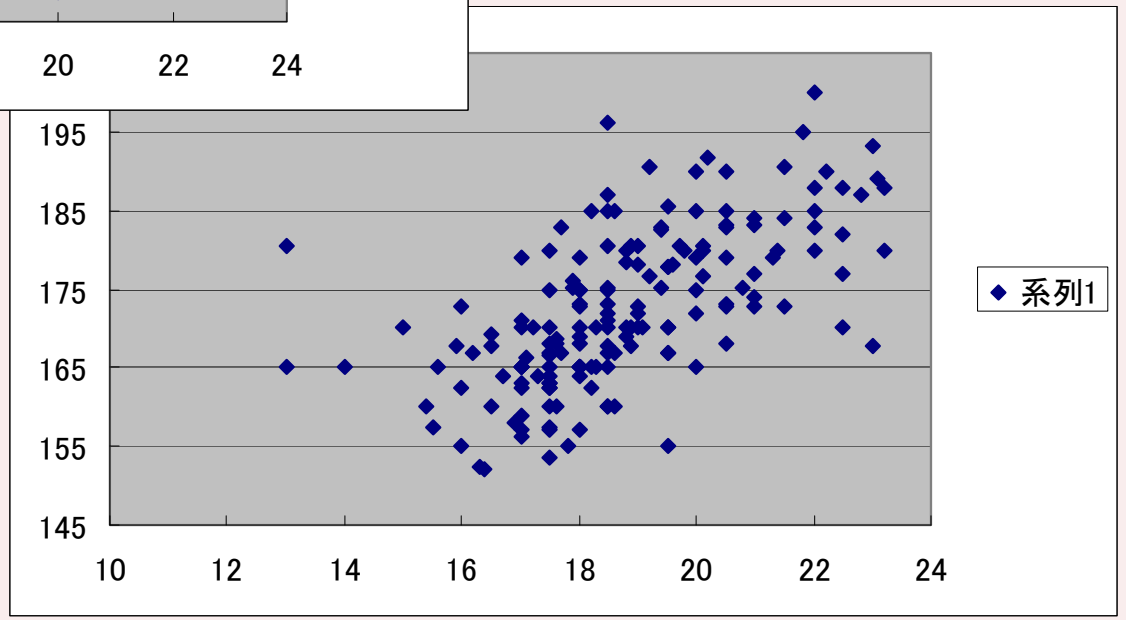
OK　キャンセル

「関係性」の強さとは？

- 一方が増えるとき他方も増える（減る）という関係が，厳密になればなるほど関係性は強く，一方が増えても他方があまり増えない（減らない）とき，関係性は弱いと言える．

関係性がある場合とない場合



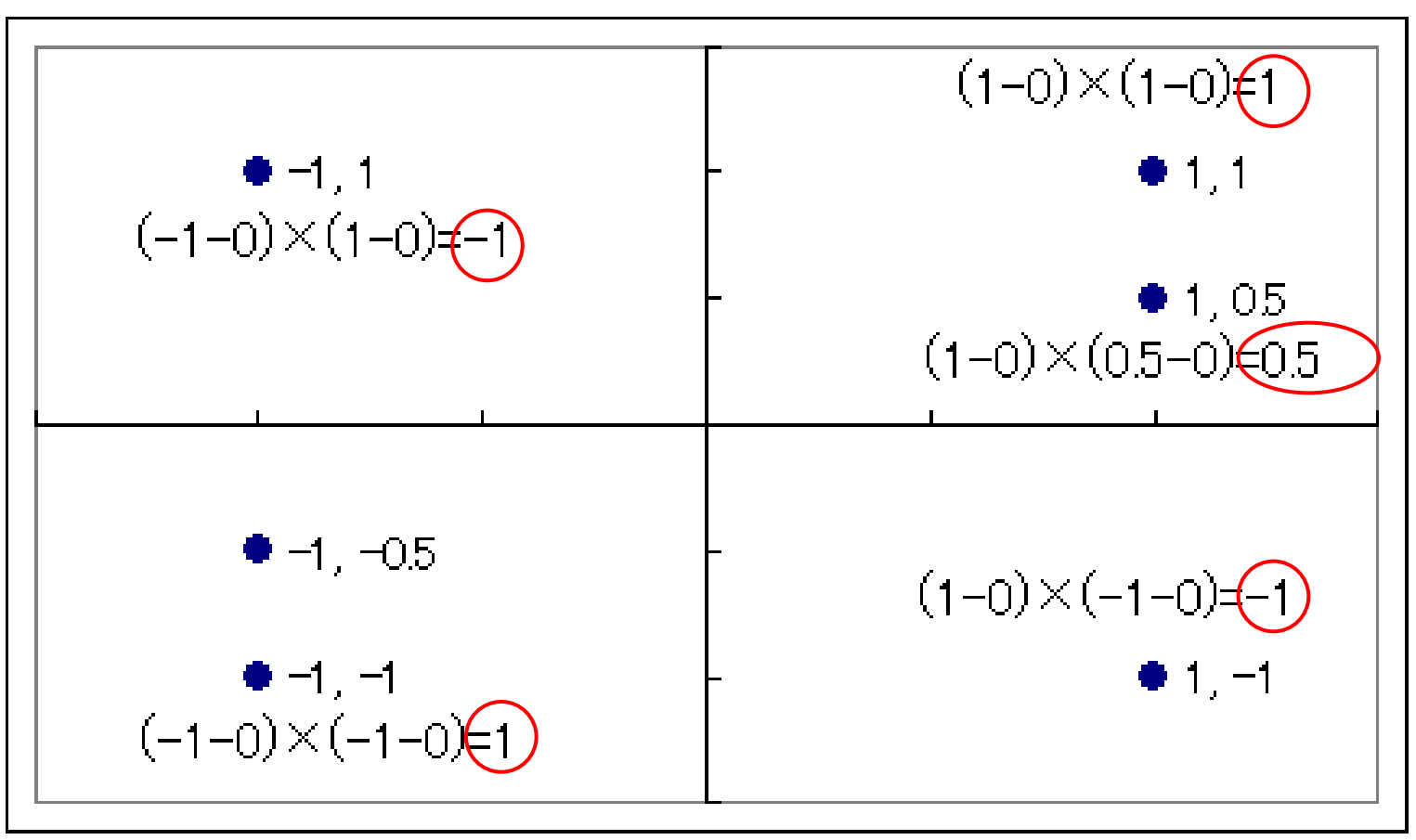


共分散

- 一方の値が大きくなると他方も大きく（小さく）なるというような，関係性の指標に共分散がある．

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

共分散の意味



- 関係性が強いときに，$(x_i - \bar{x})(y_i - \bar{y})$ の値の絶対値も大きくなる．

ピアソンの積率相関係数

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

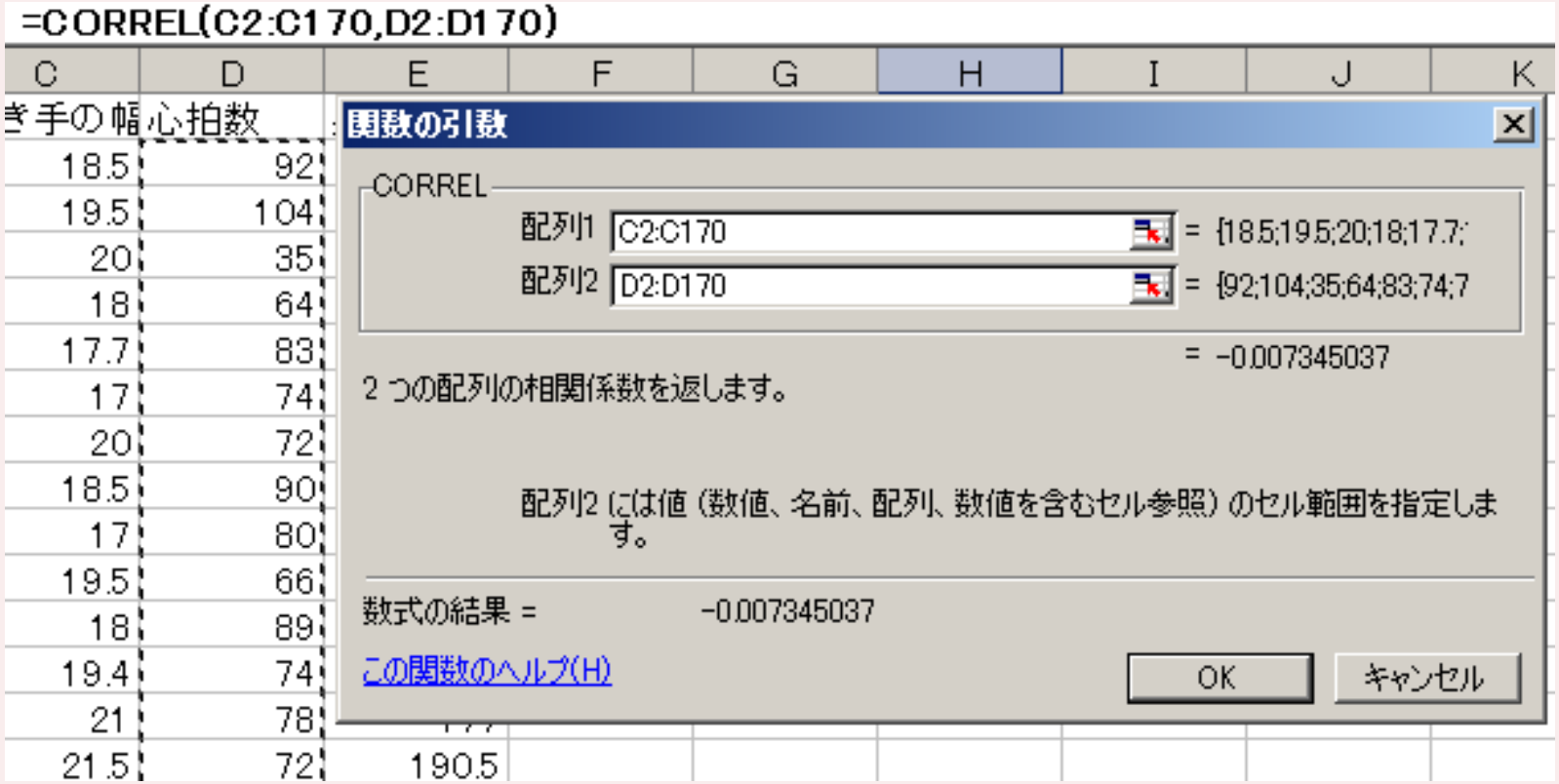
$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

- 共分散を双方の標準偏差で割ることで，相関係数の範囲は-1から1の間に調整される．

ピアソンの積率相関係数

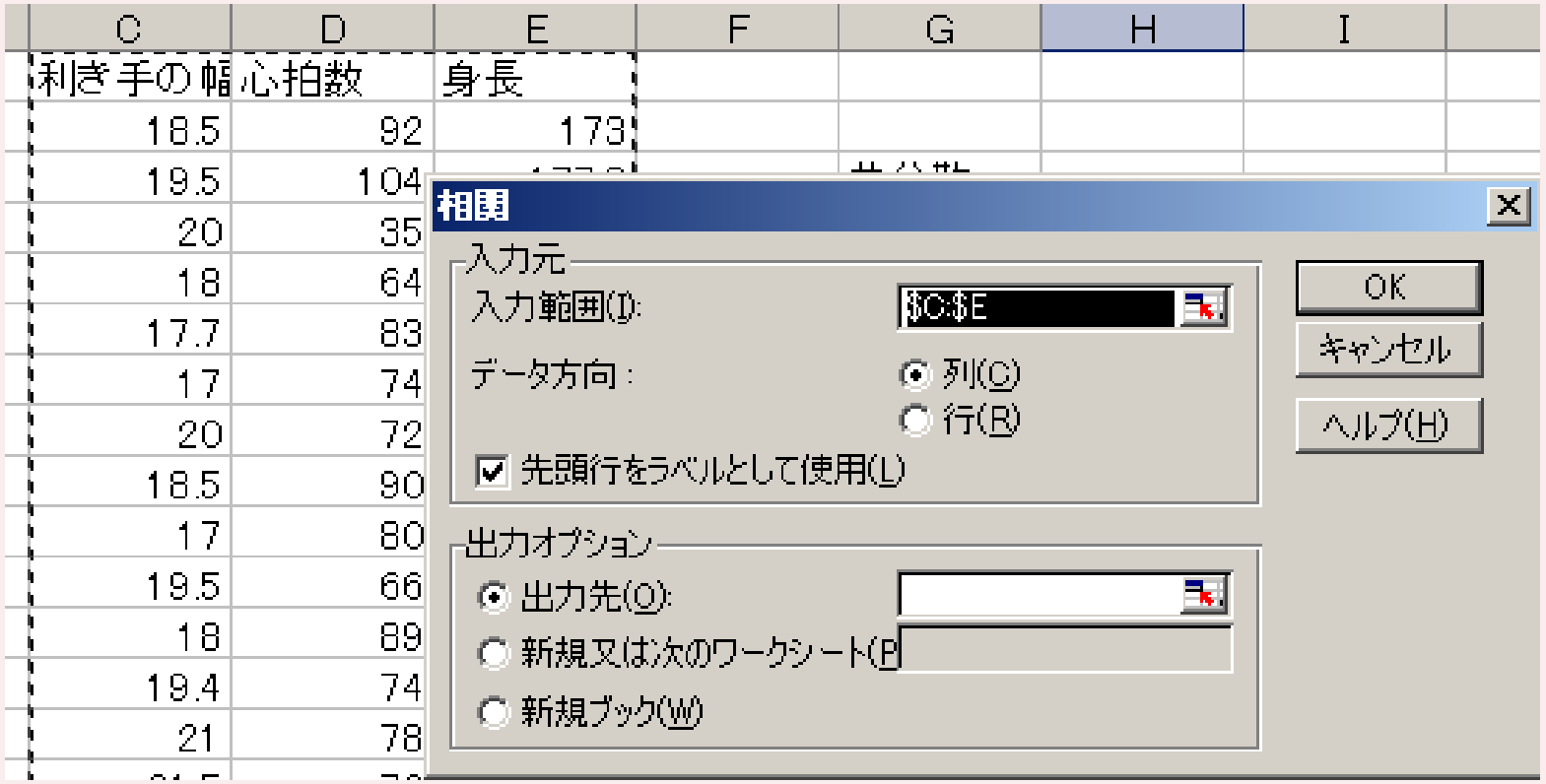
- 相関係数は，範囲が $-1 \leq \rho \leq 1$ であり，正に大きければ強い正の相関（一方が大きい値をとる場合，他方も大きい値を取りやすい），負に大きければ強い負の相関（一方が大きい値をとる場合，他方は小さい値を取りやすい）である．0に近い場合には，相関がないと考える（一方の値の大きさは他方に影響しない）．

Excelによる相関係数の算出（１）



- 関数を用いても求められるが，分析ツールを用いてもいい．

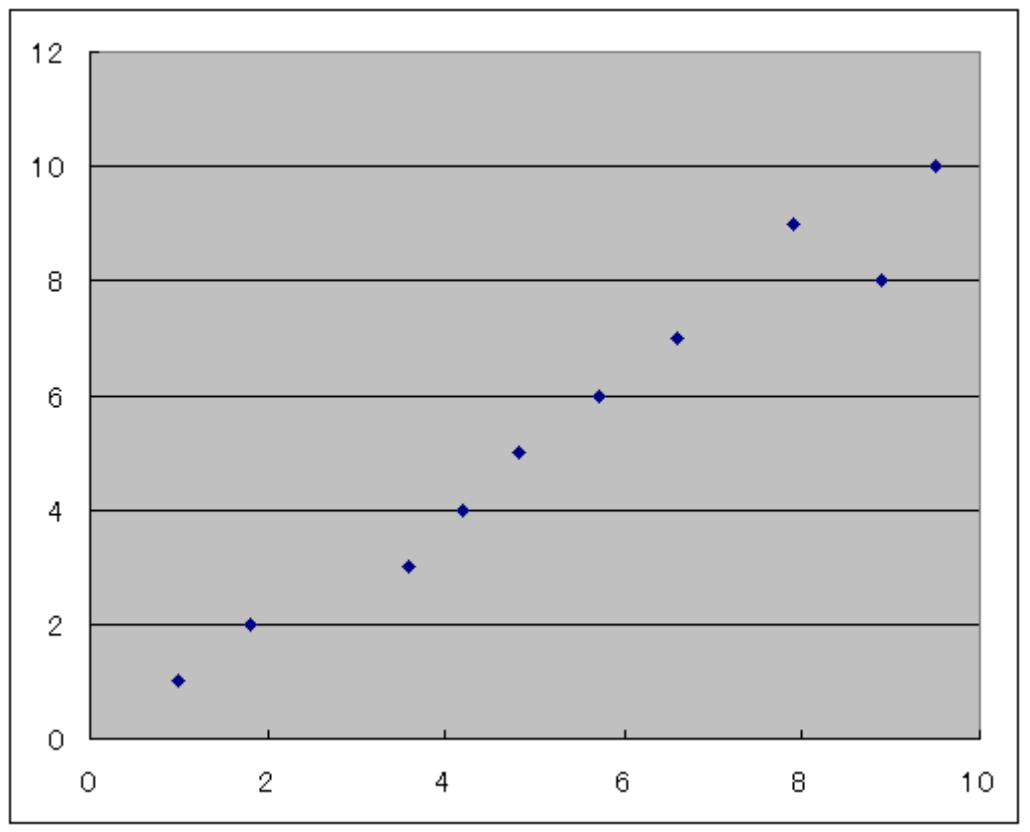
Excelによる相関係数の算出（２）



C	D	E
利き手の幅	心拍数	身長
18.5	92	173
19.5	104	
20	35	
18	64	
17.7	83	
17	74	
20	72	
18.5	90	
17	80	
19.5	66	
18	89	
19.4	74	
21	78	

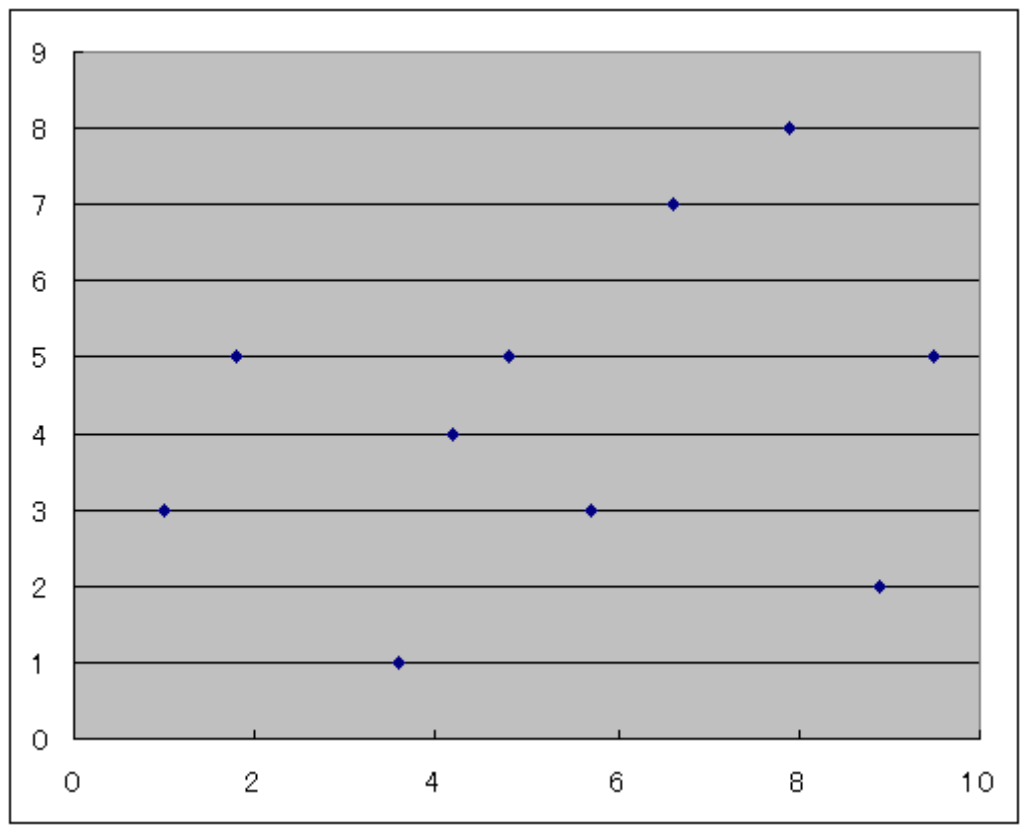
- 分析ツールは，３変数以上の相関を同時に求めるときに有用．尚，共分散を求める場合にも，関数，分析ツール双方が使用可能で，使用法も相関を求める場合とほぼ同じである．

強い正の相関



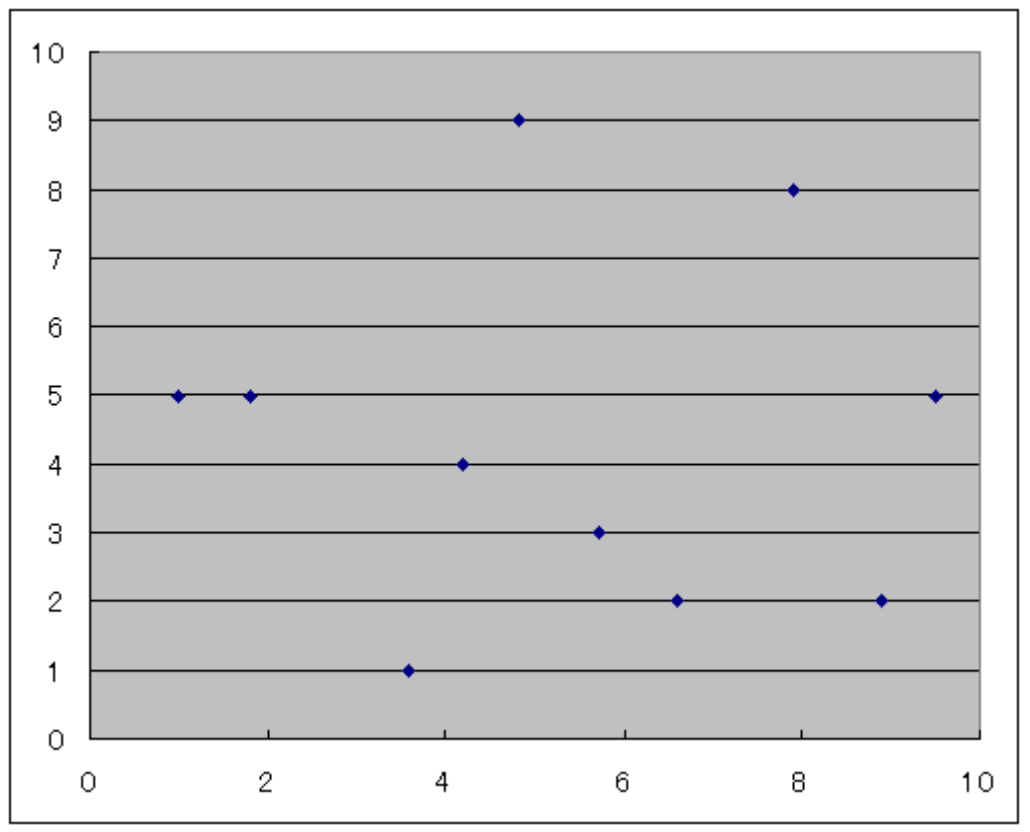
$$\rho = 0.98$$

弱い正の相関



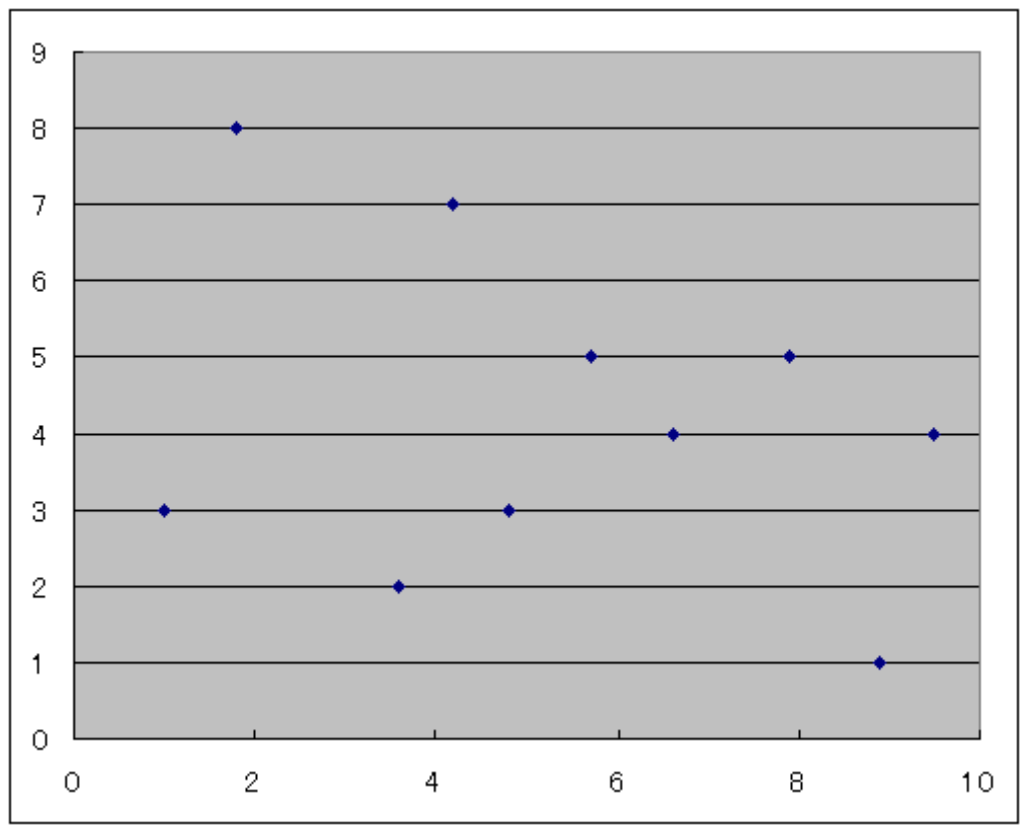
$$\rho = 0.29$$

無相関



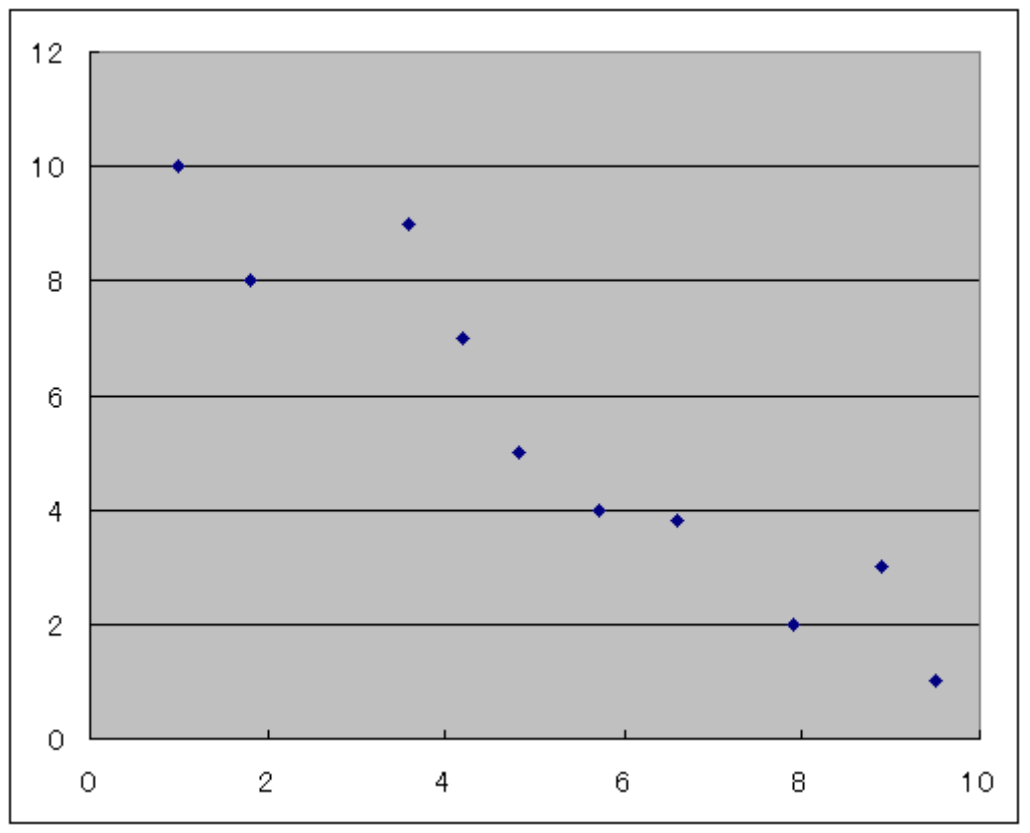
$$\rho = 0.02$$

弱い負の相関



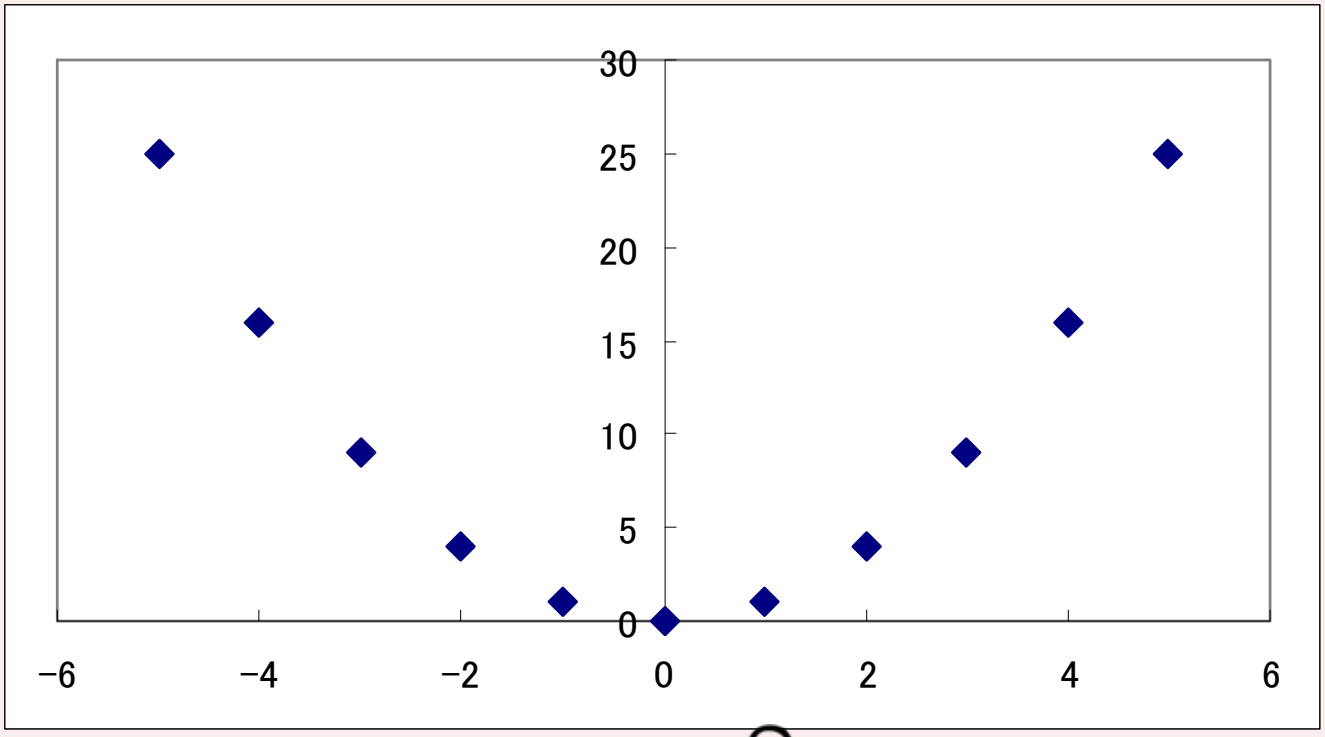
$$\rho = -0.31$$

強い負の相関



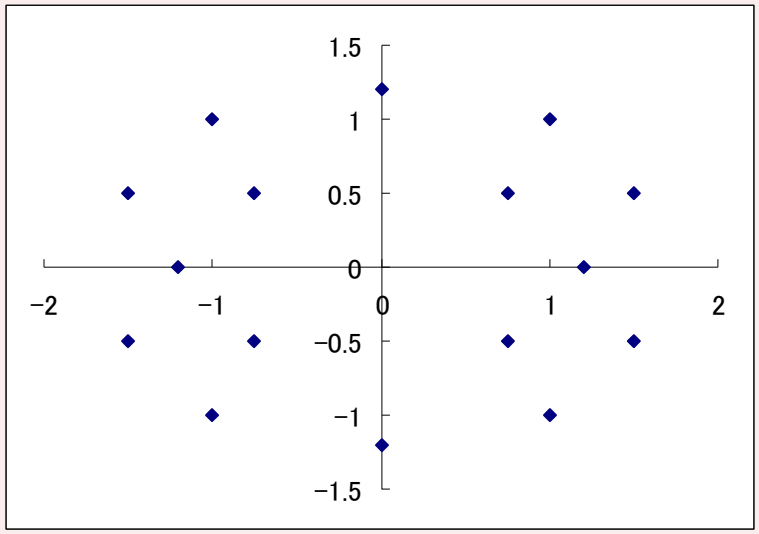
$$\rho = -0.95$$

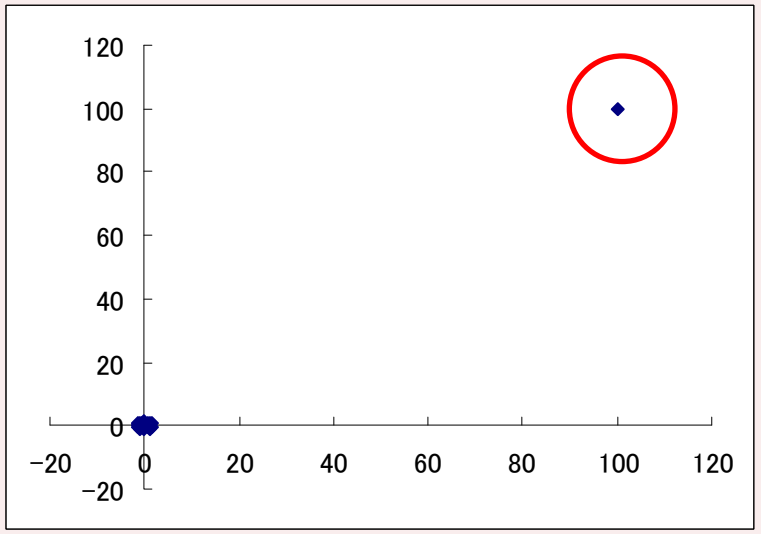
相関係数の注意点



- このデータには，$y = x^2$ という明確な関連性があるが，相関係数を算出すると $\rho = 0$ となる．この問題は，相関係数が直線的（一次関数的）な関係性しか見出せないため起こる．

相関係数の注意点





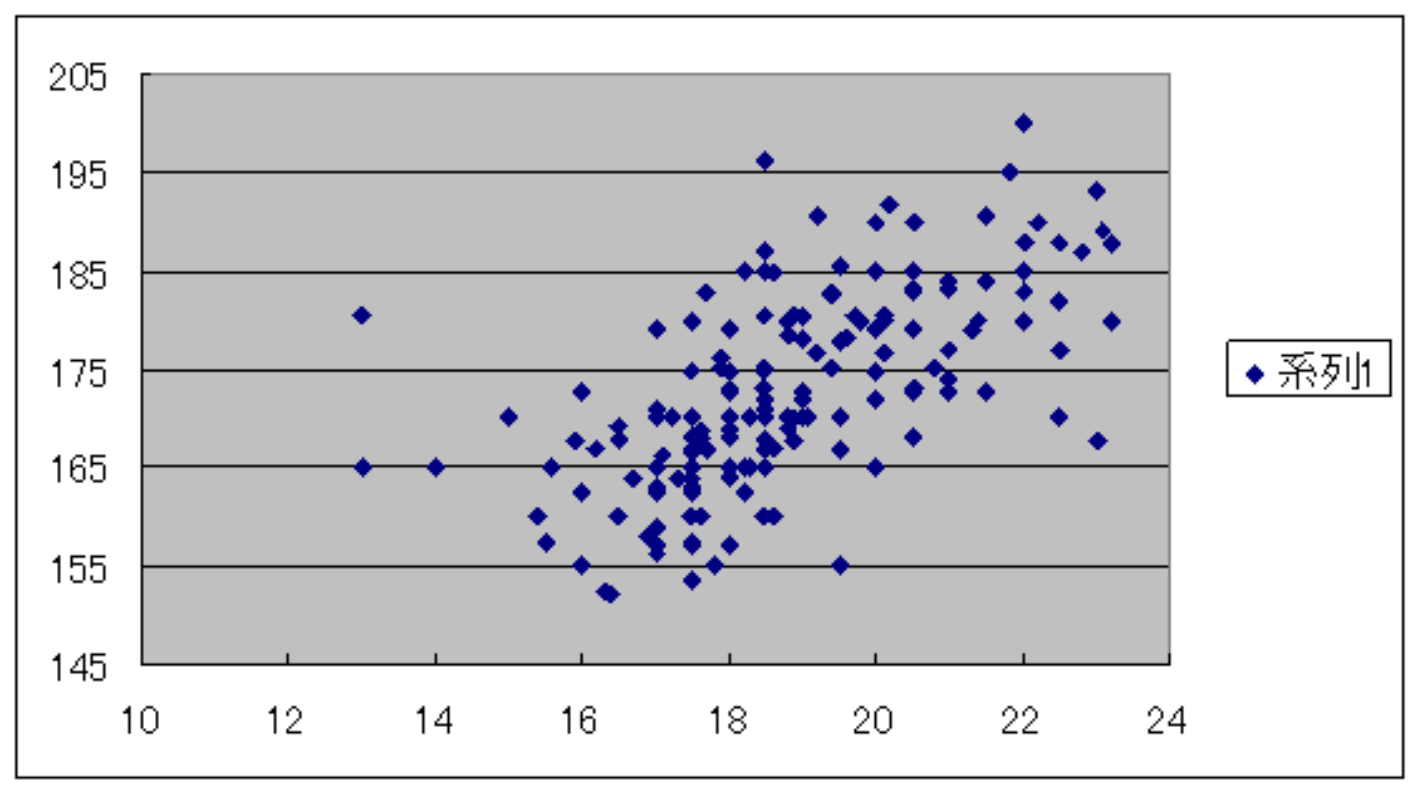
$\rho = 0$

$\rho = 0.999$

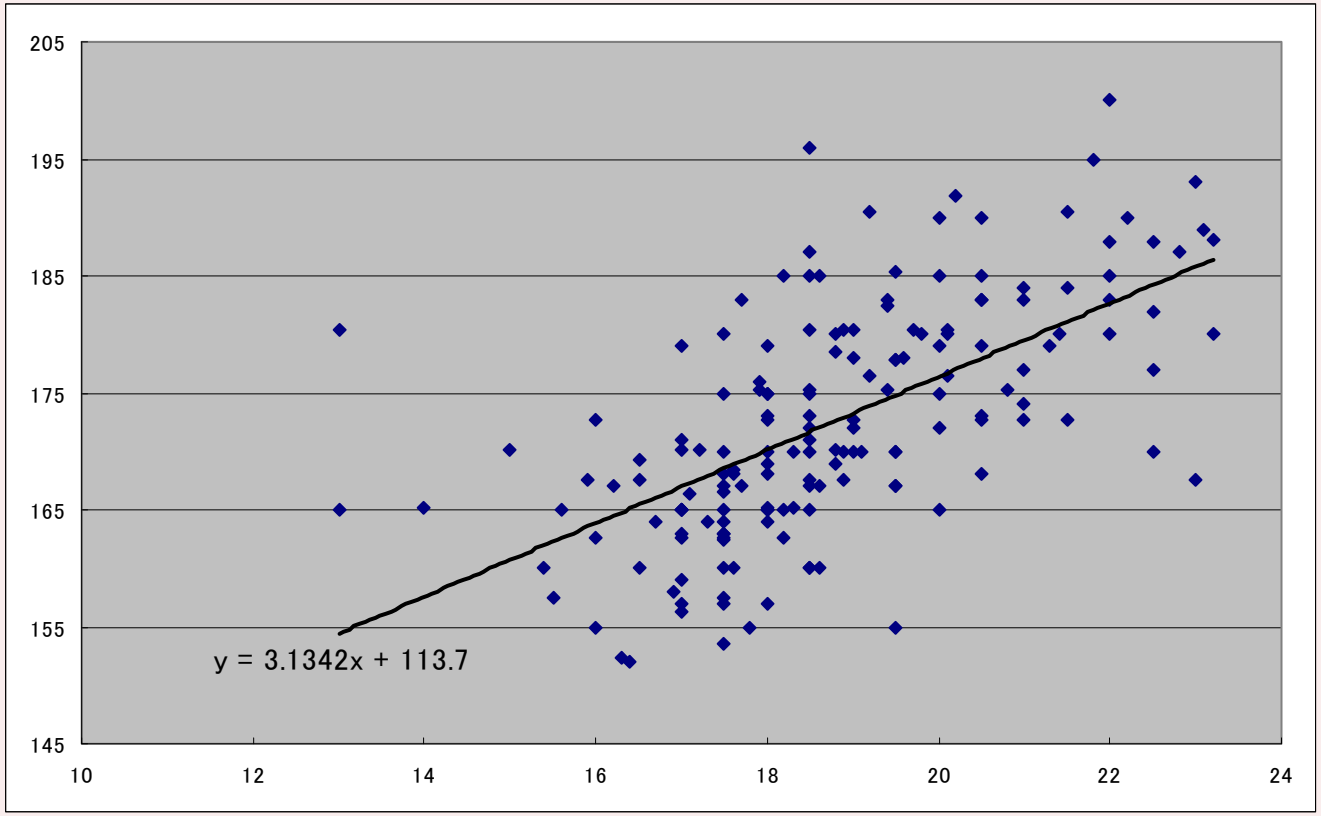
- 左の散布図のデータでは，相関は0となるが，このデータに外れ値を一箇所加えるだけで，相関はほぼ1となってしまう．このように，相関係数は外れ値に弱い点に注意が必要である．

回帰分析

- 身長を，利き手の幅から予測できないか？

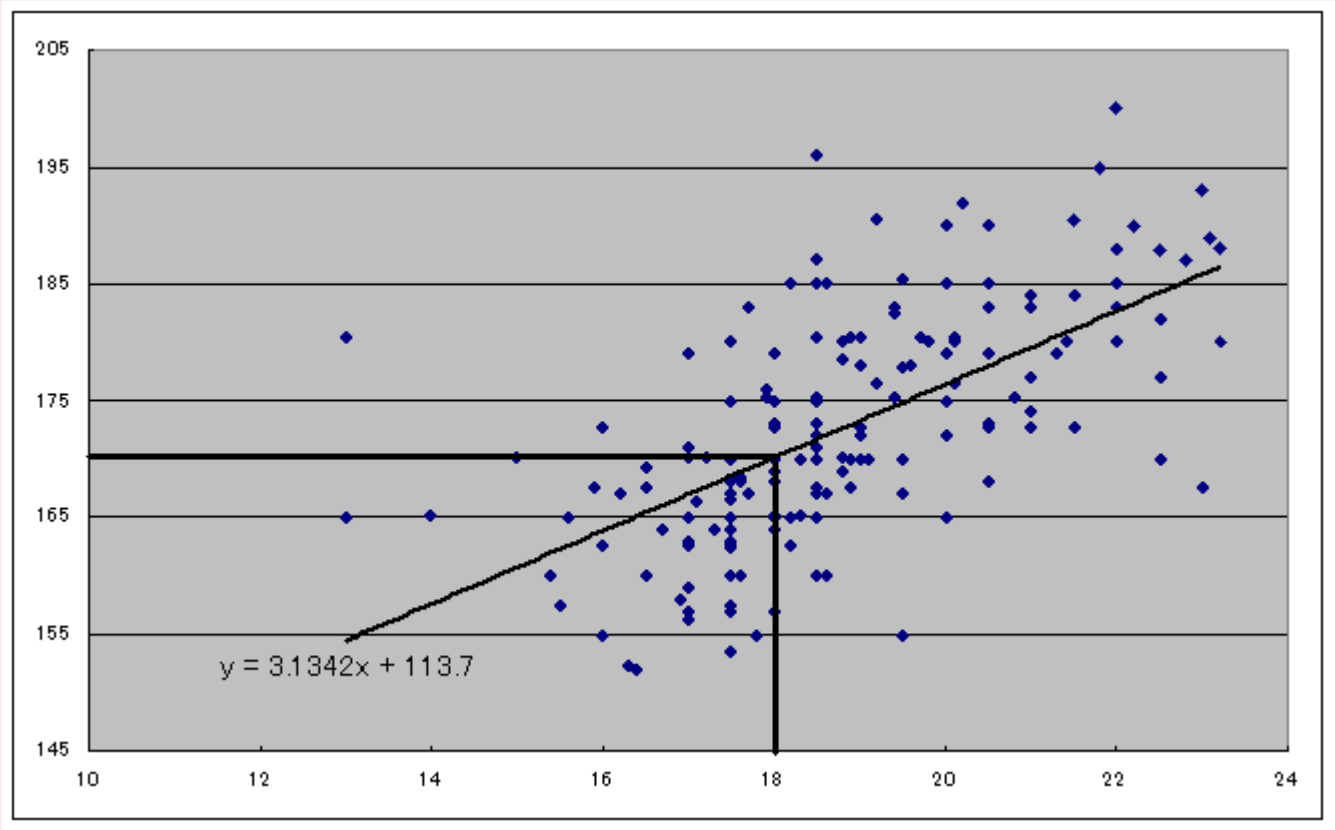


回帰分析のイメージ



回帰分析

- この場合，利き手の幅が18cmなら，身長は170.1cm程度と推測される．



回帰分析

- ある変数（予測変数・説明変数）の値から他の変数（目的変数・基準変数）の値を予測・説明する分析を回帰分析（regression analysis）と呼ぶ．特に，一次関数を用いて予測・説明を行う場合，線形回帰と呼ぶ．また，予測変数が１つの場合を単回帰分析，複数ある場合を重回帰分析と呼ぶ．
  - 「利き手の幅」から「身長」を予測・説明．
    →単回帰分析
  - 「利き手の幅」と「心拍数」から，「身長」を予測・説明．
    →重回帰分析

単回帰分析

- 単回帰分析では，以下のようなモデル式をまず考える．

$$y_i = \alpha x_i + \beta + e_i$$

- この式は，前々回説明した一元配置の分散分析と記号は少々異なるが，ほぼ同じ形である．分散分析と回帰分析は，線形モデルとして，数学的性質を同じくする．
- ここで，y_iは目的変数（身長など），x_iは予測変数（利き手の幅など）であり，α, β はそれぞれ一次関数における傾き，切片であり，e_i は誤差である．

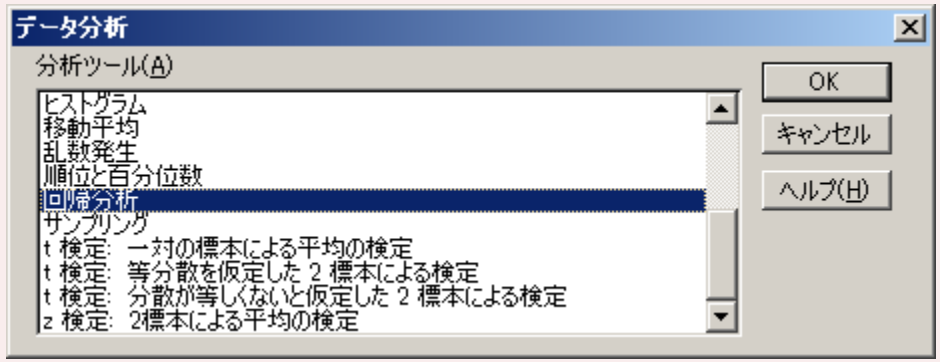
単回帰分析

$$y_i = \alpha x_i + \beta + e_i$$

予測部分　　誤差

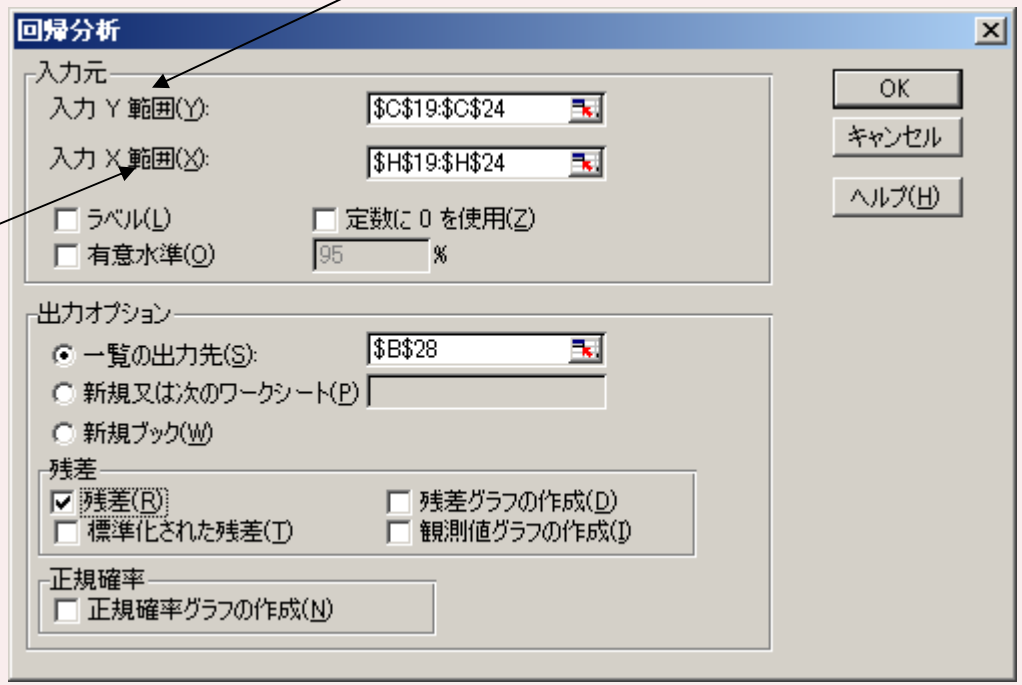
- この式は，x_iを利用してy_iを，$\hat{y}_i = \alpha x_i + \beta$という形で予測，説明する事を意味している．
- しかし，予測には誤差が必ず含まれる（同じ利き手の幅であったとしても，同じ身長になるとは限らない）その誤差がe_iで表現される．
- このα, βの値をデータから推定することで，予測式を完成させる．

Excelによる回帰分析



目的変数

予測変数



単回帰分析の出力

	係数
切片	113.7
利き手の幅	3.134201

係数の推定値

$$\hat{y} = 3.134201x + 113.7$$

重回帰分析の出力

	係数
切片	118.8849
利き手の幅	3.131219
心拍数	-0.06931

係数の推定値

$$\hat{y} = 3.131219x_1 - 0.06931x_2 + 118.8849$$

実習

- chp05_c.xlsのデータを用い，日経平均，ANA，JALそれぞれの株価について相関係数を求めよ．
- chp10_a.xlsのデータを用い，年齢（実際には月齢）を横軸，身長を縦軸に置いた散布図を描き，相関係数を求めよ．また，月齢から身長を予測する回帰分析モデルを作成せよ．また，月齢が18ヶ月の時の身長の予測値を求めよ．