

必要なサンプル数（サンプルサイズ）

$$n = \frac{N}{\left(\frac{E}{1.96}\right)^2 \times \frac{N-1}{0.25} + 1}$$

「サンプルサイズ」⇒ n : 必要なサンプル数

N : 母集団の数

E : 誤差の範囲（普通は0.05）



ランダムサンプリング
で取ってくる数ね！

ちなみに

前回の母集団の数（12000、5500）

で必要なサンプル数を計算してみたら

	誤差5%	誤差1%
5000	359.14	372.28
12000	3497.45	5336.79

必要なサンプル数（サンプルサイズ）

$$n = \frac{N}{\left(\frac{E}{1.96}\right)^2 \times \frac{N-1}{0.25} + 1}$$

「サンプルサイズ」⇒ n : 必要なサンプル数

N : 母集団の数

E : 誤差の範囲（普通は0.05）

「誤差」

結果が本当の値から

ある程度ズれる

ことを見込む



例えば (0.05) など...

一番やったらダメな間違い！！

5%」正しい！

E : 誤差の範囲（普通は0.05）

簡単に言うと

一部を調べた結果

AaaAA

全員を調べた結果

AAAAA

「全員を調べた結果」と「一部を調べた結果」

は少しズれる

誤差

普通（一般的に）よく使われるのは

本当の値から

「0.05」

「 $\pm 5\%$ くらいのズレ」

までOK！

本当の値から

「0.01」

「 $\pm 1\%$ くらいのズレ」

までしか許さない！



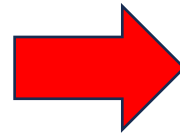
E : 誤差の範囲（普通は0.05）

サンプル

一部を調べた結果

賛成

60%



母集団

全部を調べた結果

賛成

55~65%

母集団では、だいたい
55~65%くらいかな

どっちにするかは、

どれだけ正確なデータが欲しいかで変えればいい！

誤差の範囲

「0.05 (5%)」

必要サンプル数

「少ない」



誤差の範囲

「0.01 (1%)」

必要サンプル数

「多い」

必要なサンプル数 (サンプルサイズ)

だいたいどれくらいの
サンプル数がある程度正確なの？

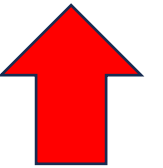


一般的に言われてるのは
だいたい 最低でも 400

母集団の数	誤差の範囲			
	1%	3%	5%	10%
500	475	340	217	80
1000	905	516	277	87
3000	2286	787	340	93
5000	3288	879	356	94
10000	4899	964	370	95
100000	8762	1055	382	96
1000000	9512	1066	384	96

10%の誤差が出てもいいなら

母集団の数にかかわらず100件取れば十分！



どんどん統計の本体へ近づいてるよ

「相関係数」



「相関係数」

相関係数：2種類のデータの関係性の強さを
「-1 から +1」 の値で表した数
「 r 」で表されることが多い



「相関の強さ」

$$0.7 \leq r \leq 1.0$$

$$0.4 \leq r \leq 0.7$$

$$0.2 \leq r \leq 0.4$$

$$-0.2 \leq r \leq 0.2$$

$$-0.4 \leq r \leq -0.2$$

$$-0.7 \leq r \leq -0.4$$

$$-1.0 \leq r \leq -0.7$$

強い正の相関

正の相関

弱い正の相関

相関なし

弱い負の相関

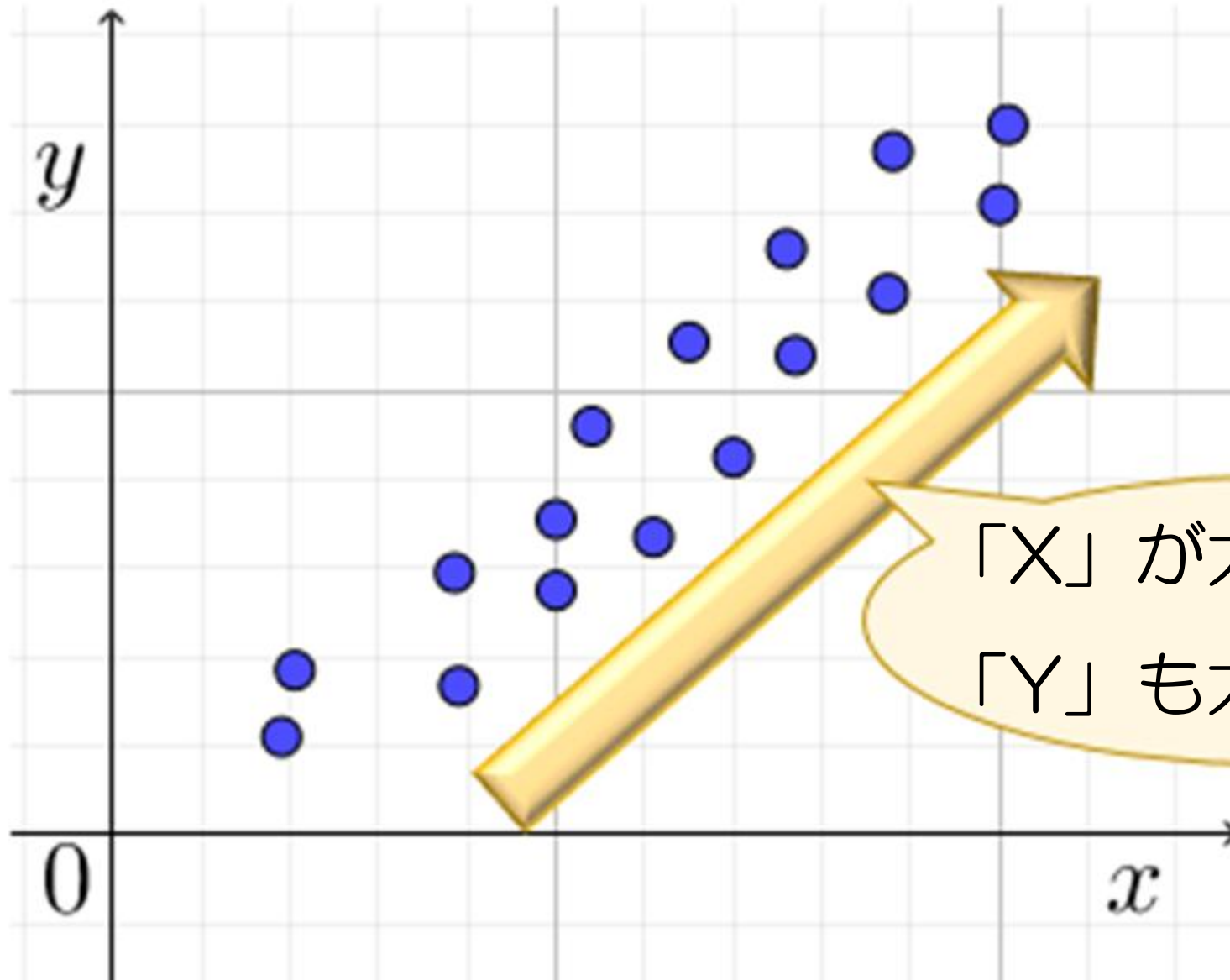
負の相関

強い負の相関

「X」が大きくなると
「Y」も大きくなる

「X」が大きくなると
「Y」は小さくなる

例えば「相関係数：0.94」



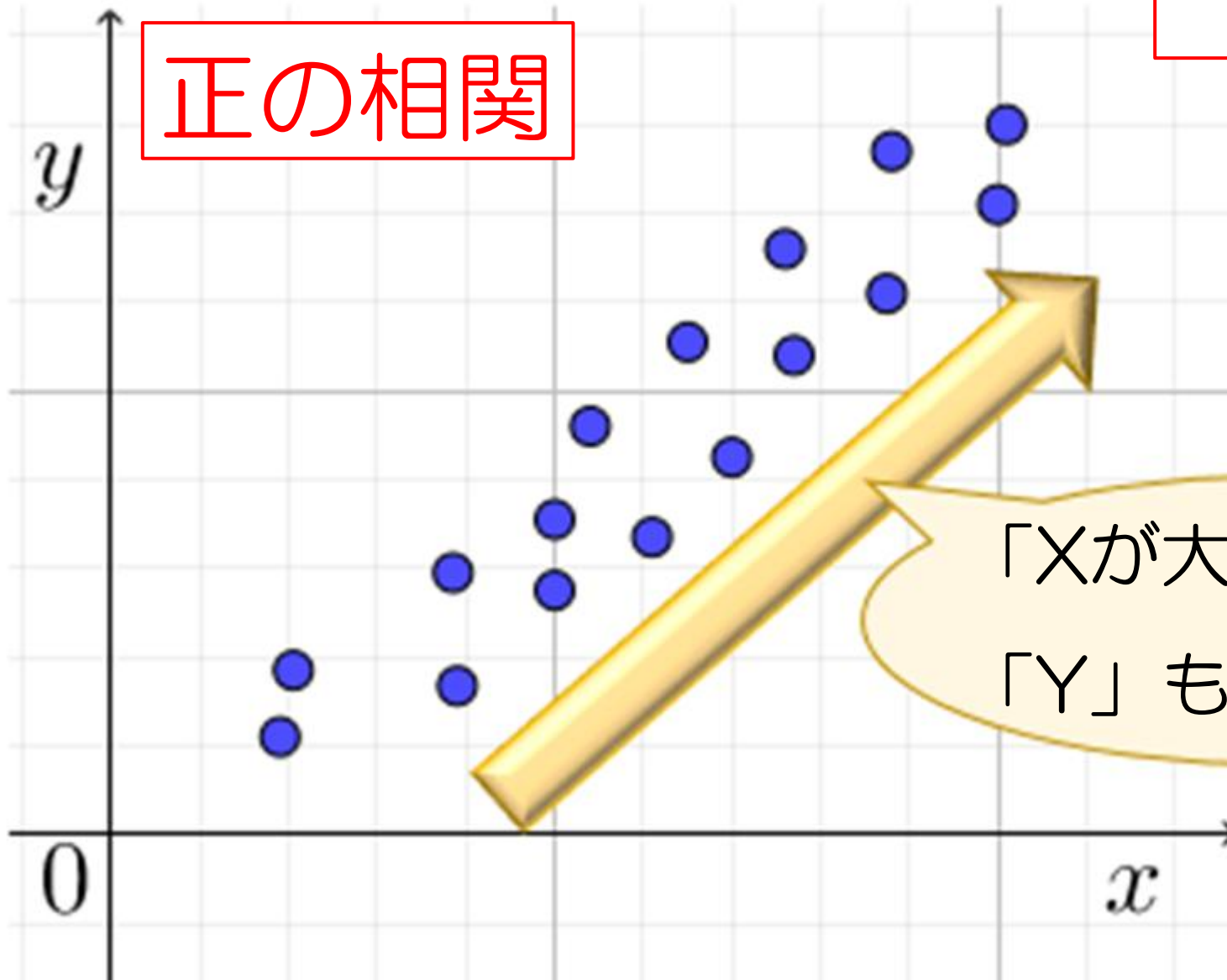
「X」が大きくなると
「Y」も大きくなる



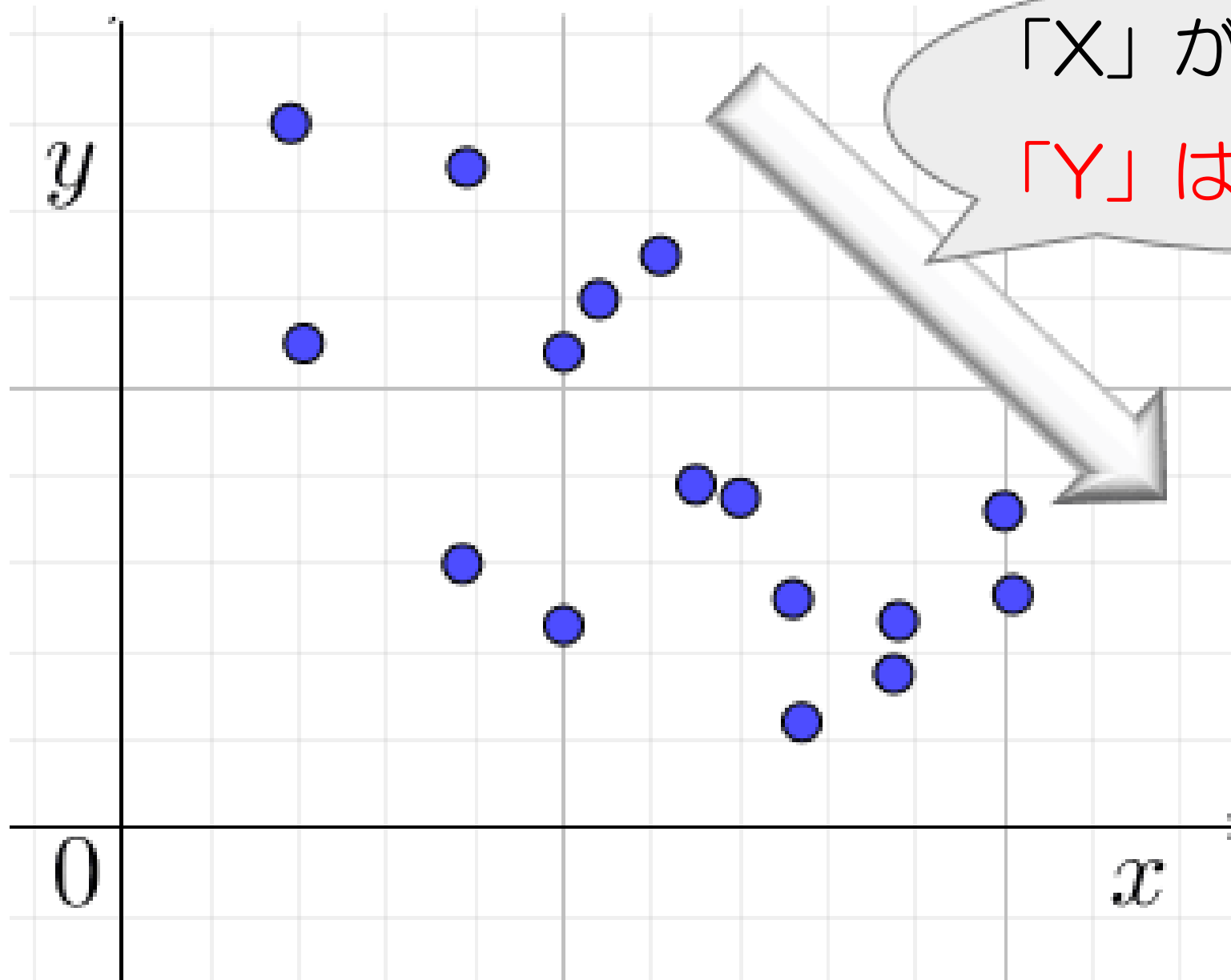
例えば「相関係数：0.94」

身長が大きくなると
体重も増える

正の相関



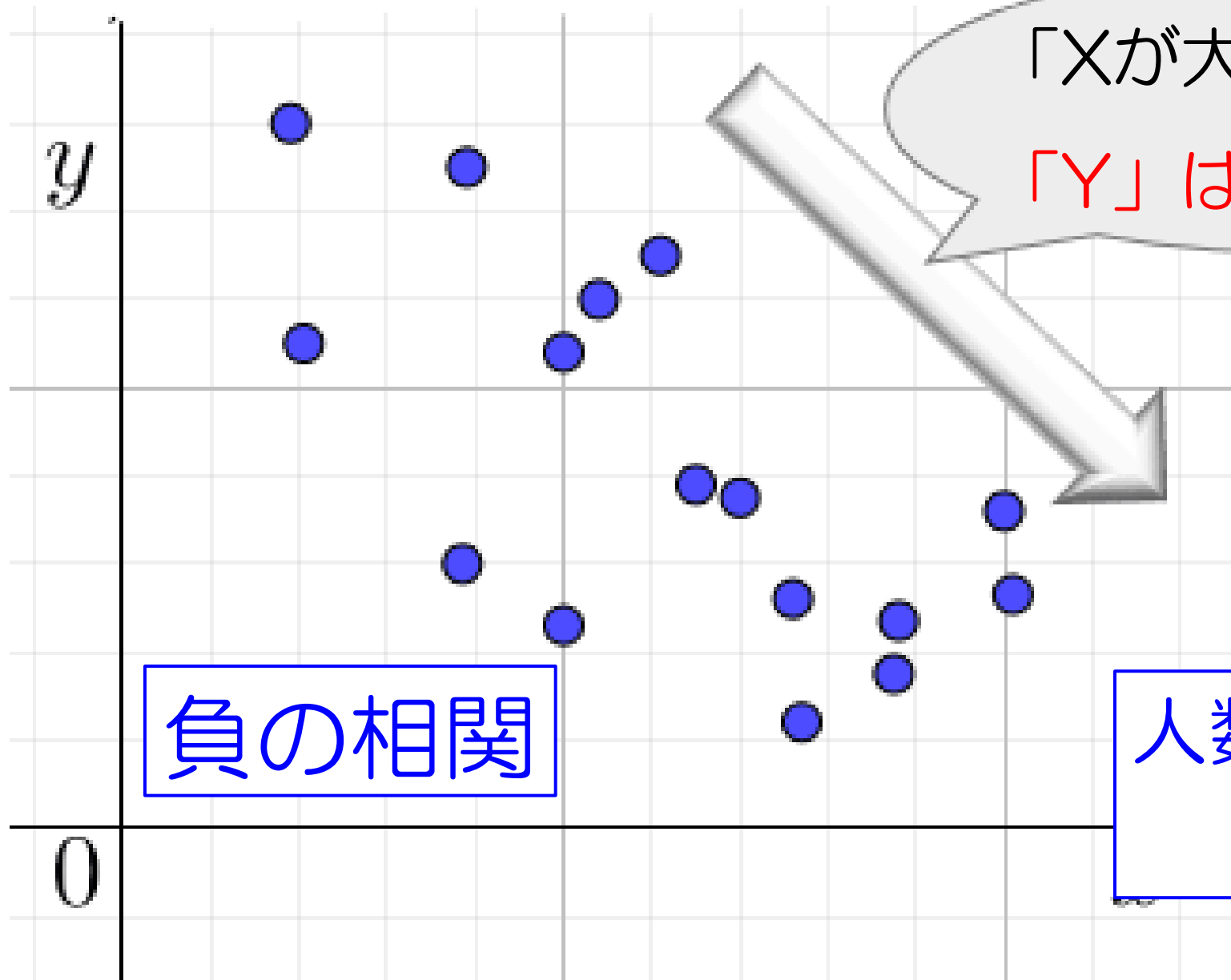
例えば「相関係数：-0.76」



「X」が大きくなると
「Y」は小さくなる



例えば「相関係数： -0.76 」



「Xが大きくなると
「Y」は小さくなる

負の相関



人数が増えると
分け前が減る

相関係数には弱点が3つある

- 1 サンプル数が少ないと、正確じゃない
- 2 外れ値があると、正確ではないかも
- 3 第3の要因があることがある

1 サンプル数が少ないと、正確じゃない

前回は「全国レベル」やったけど
町内とかのレベルのデータでは、

全く違った結果が得られるため

ほとんど意味が無い

ということは、大量のデータが無いと使えない！



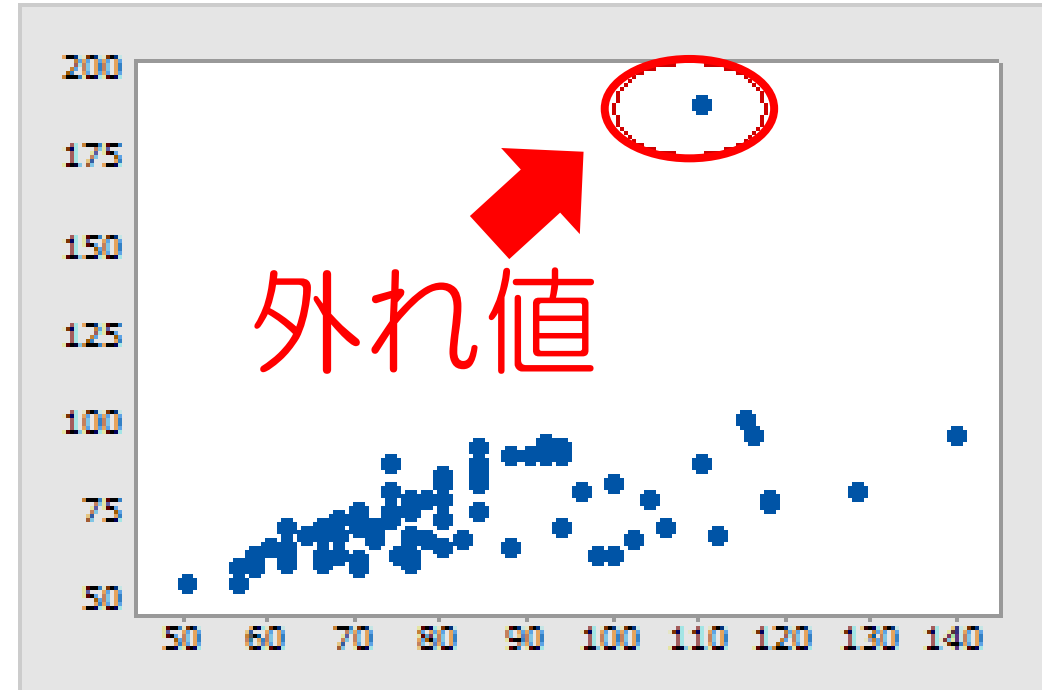
2 外れ値があると、正確ではないかも

「データが**きちんと分布している**」ことが前提

「外れ値」があるデータからの
相関係数は、信用度が下がる



正確な相関係数であることも
あるけど**判断できない**



3 第3の要因があることがある

相関関係があるからといって

因果関係があるとは限らない！

「身長が高いほうがテストの点がいい」

という相関関係がでた

本当？



3 第3の要因があることがある

「身長が高いほうがテストの点がいい」

例えば、小学校1年生と高校生では
高校生の方がでかいし、勉強もできる

身長とテストの点数以外に

「年齢」という要因があった！

(第3の要因)



「相関関係」と「因果関係」

相関関係：一方が増加すると他方も変動する

因果関係：原因とそれによって生まれた結果



雨に濡れたまま寝たら、
風邪をひいた



相関係数の求め方

$$r = \frac{S_{xy}}{S_x \times S_y} = \frac{\frac{1}{n} \sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=0}^n (x_i - \bar{x})^2} \times \sqrt{\frac{1}{n} \sum_{i=0}^n (y_i - \bar{y})^2}}$$

r : x と y の相関関数

S_{xy} : x と y の共分散

S_x : x の標準偏差

S_y : y の標準偏差

わけわからん！



Excelを使いましょう！

標準偏差：「STDEV.P」関数

共分散：「COVAR」関数

相関係数：「CORREL」関数

予測つくかもやけど、
もっと楽ながあります！



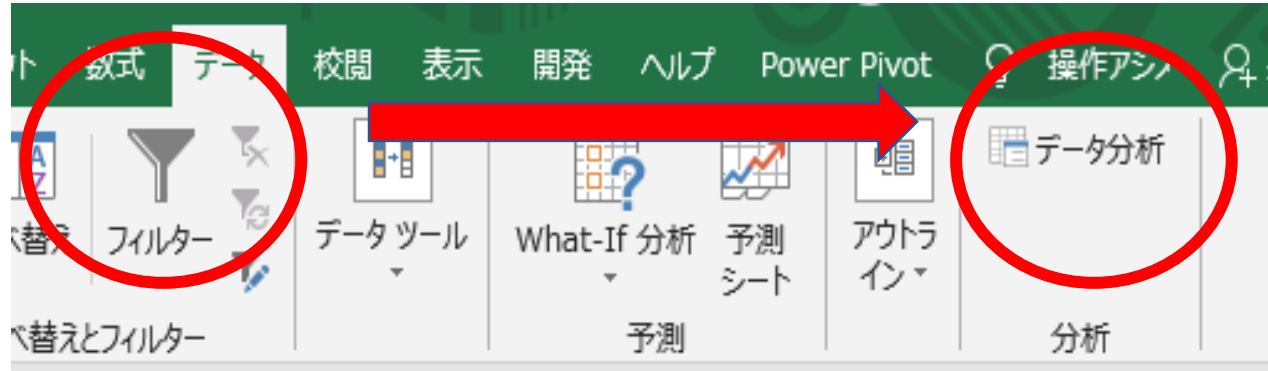
「標本の「相関係数」求め方」



Excelの「データ」タブの中にある

「データ分析」 → 「相関」

標本の「相関係数」求め方



「データ分析」

⇒ 「相関」



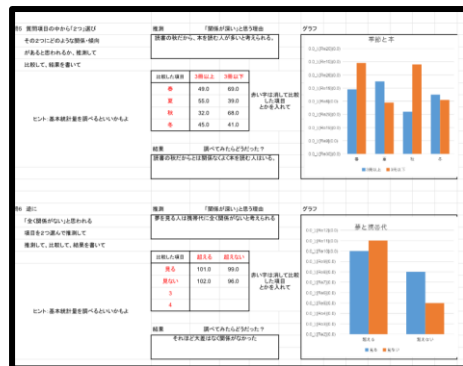
今回は全範囲選択OK！
やし、一瞬でできる

提出0

1 回目の授業でやった アンケート結果の「相関」調べてみよう！！

問5 項目の中から質問を「2つ」選び
その2つにどのような関係・傾向
があると思われるか、推測して

- ・「平均値」
- ・「度数分布表」
- ・「グラフ」(何でもOK)
- ・「標準偏差」 とかも忘れずに！



自分が

「関係がある」と判断したものが
あってたのか、間違ってたのか

数値になるから面白い！

問6 逆に

「全く関係がない」と思われる
項目を2つ選んで推測して

- ・「平均値」
- ・「度数分布表」
- ・「グラフ」(何でもOK)
- ・「標準偏差」 とかも忘れずに！

提出2

「都道府県別 医療データ」があるから

その中から

問題1 正の相関関係にあるもの

問題2 負の相関関係にあるもの

を見つけてみよう！

当たり前のものは、アカンで！

例えば、人口が多いと死亡者数が多いとか！

提出 3

「都道府県別 警察データ」があるから

その中から

問題 1 正の相関関係にあるもの

問題 2 負の相関関係にあるもの

を見つけてみよう！

大学生が多いと〇〇が多いとか、

やくざが多いと〇〇が少ないとか、面白いのいっぱいある

Excelを使いましょう！

標準偏差：「STDEV.P」関数

共分散：「COVAR」関数

相関係数：「CORREL」関数