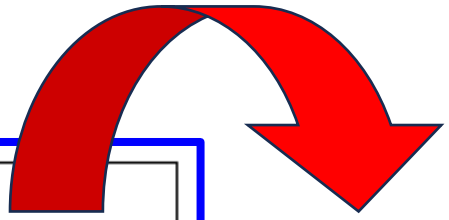
「収入は？」



サンプル数	３人だけの標準偏差
20	0.094634
100	0.044969
500	0.021602
2000	0.005907

サンプ	標準偏差
20	0.093073
10	0.044653
50	0.023658
200	0.005892

多少の違いはあるけど、「推測統計」

ある程度母集団の傾向をとらえてる

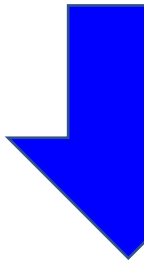
「支出は？」

サンプル数	Aさん	Bさん	Cくん	標準偏差
20	0.875	0.917	0.974	0.040730
100	1.048	1.021	1.052	0.013877
500	0.985	1.006	0.989	0.009038
2000	0.987	1.01	0.998	0.009448

ほぼ正確だけど、わずかに違う！

これが「標本誤差」

今回は「１に近いかどうか」だけで判断した



だいたいどれくらいの

サンプル数があればいいの？

「正しく反映できてる」って言っていいの？

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

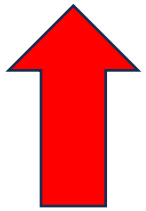
E ：誤差の範囲（普通は0.05）

E：誤差の範囲（普通は0.05）

普通（一般的に）よく使われるのは

「0.05」⇒　変なデータが出る確率は「５％」

「0.01」⇒　変なデータが出る確率は「１％」



言い換えると、

母集団を正確に反映していないデータが出る確率

E：誤差の範囲（普通は0.05）

誤差（とんでもなく外れてる値）が

出る確率は「５％」ってこと

今回のデ

一番やったらダメな間違い！！

「９５％」の確率で正しい！

E：誤差の範囲（普通は0.05）

これが正解！

「今回のデータが正しい」と仮定したら…

変なデータが出る確率は「５％」

たまたまそんなに低い確率は考えにくいから

「今回のデータが正しい」という仮説は

正しいと言ってもOK！

なに言ってるかわからん！

関西人に「新大阪駅についてアンケート」を取りたい

関西人全員に答えてもらえばいいんだけど（全数調査）

実際にはほぼ不可能。

その辺にいてる100人に

アンケートを取る！

その100人の中に「東京人」が数人いるかも！

⇒　母集団の意見の傾向とは違う！

誤差の範囲「0.05（5%）」でアンケートを取る

⇒　100人中５人までは東京人がいてもいい

誤差の範囲「0.01（1%）」でアンケートを取る

⇒　100人中1人までしか東京人はダメ！

その100人の中に「東京人」が数人いるかも！

⇒　母集団の意見の傾向とは違う！

誤差の範囲「0.05（5%）」でアンケートを取る

⇒　100人中５人までは東京人がいてもいい

東京人の割合が多くてもいいので

サンプルサイズは小さくなる

⇒　ただ、データの信頼性は低い

その100人の中に「東京人」が数人いるかも！

⇒　母集団の意見の傾向とは違う！

東京人の割合がめちゃ少なくないとダメなので

サンプルサイズも大きくなる！

⇒　ただ、データの信頼性は高い

誤差の範囲「0.01（1％）」でアンケートを取る

⇒　100人中1人までしか東京人はダメ！

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%）」

誤差の範囲
「0.01（1%）」

例えば

「学園祭やるか、やらないか」

誤差の範囲
「0.05（5%）」

誤差の範囲
「0.01（1%）」

アンケート結果で「95%がやる」って

出たとすれば、母集団の「90～100%」が賛成

別に多少の誤差があってもいい！

例えば

「副作用が出ない、出たか」

誤差の範囲
「0.05（5%）」

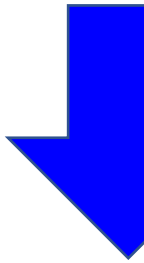
誤差の範囲
「0.01（1%）」

アンケート結果で「95%が出た」って

出たとすれば、母集団の「90～100%」が出る！

こういう場合は、誤差は少ないほうがいい

今回は「１に近いかどうか」だけで判断した



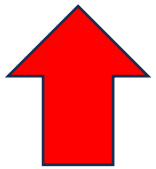
だいたいどれくらいの

サンプル数があればいいの？

「正しく反映できてる」って言っていいの？

必要なサンプル数（サンプルサイズ）

	誤差の範囲			
母集団の数	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！

逆に前回やった、

必要なサンプル数（サンプルサイズ）

「20」「100」「500」「2000」は

母集団の数が
「〇〇〇人くらい」の時に使える？

誤差の範囲「0.05」か「0.01」で固定

「相関の強さ」

$0.7 \leq r \leq 1.0$	強い正の相関
$0.4 \leq r \leq 0.7$	正の相関
$0.2 \leq r \leq 0.4$	弱い正の相関
$-0.2 \leq r \leq 0.2$	相関なし
$-0.4 \leq r \leq -0.2$	弱い負の相関
$-0.7 \leq r \leq -0.4$	負の相関
$-1.0 \leq r \leq -0.7$	強い負の相関

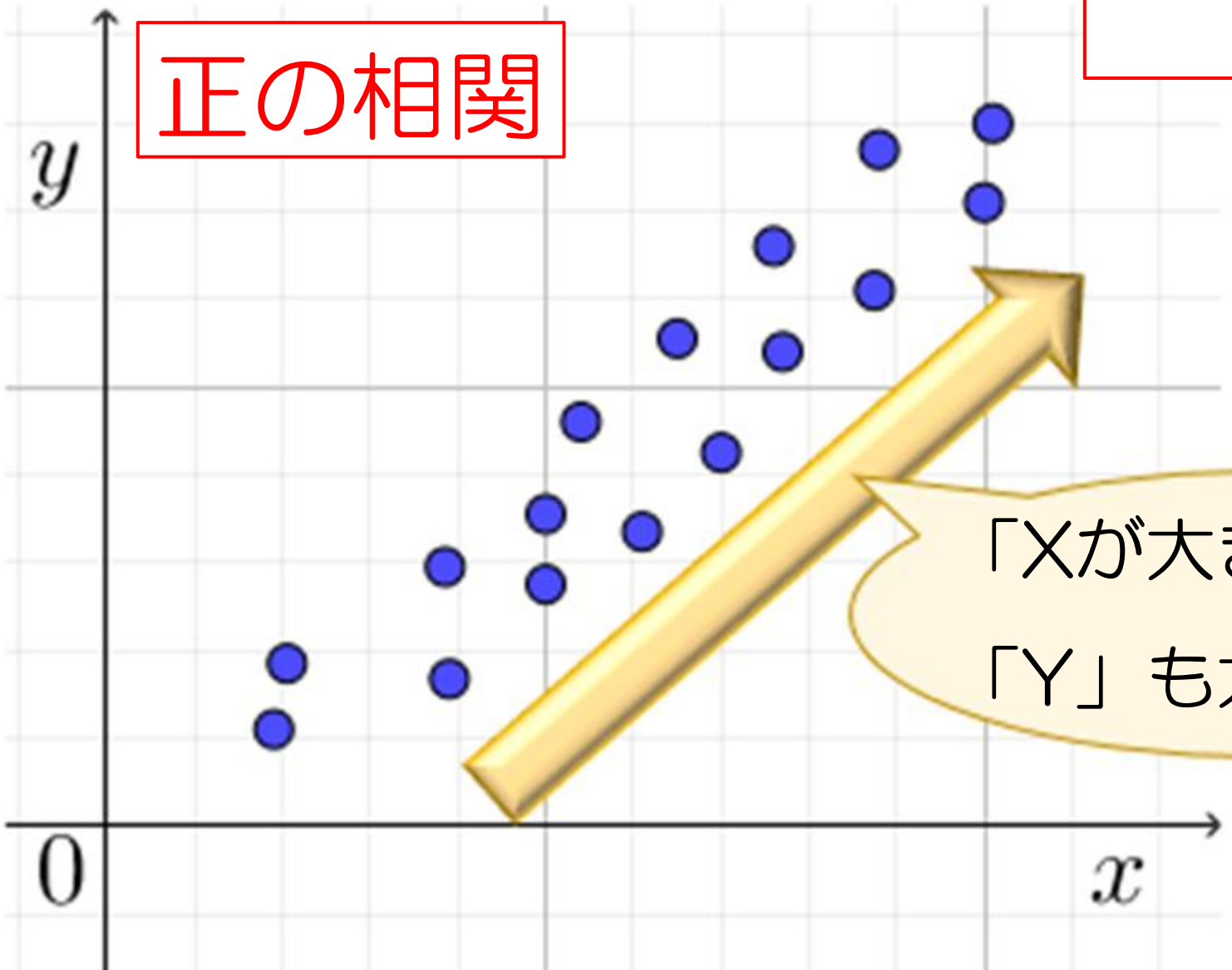
「X」が大きくなると
「Y」も大きくなる

「X」が大きくなると
「Y」は小さくなる

「相関係数：0.94」

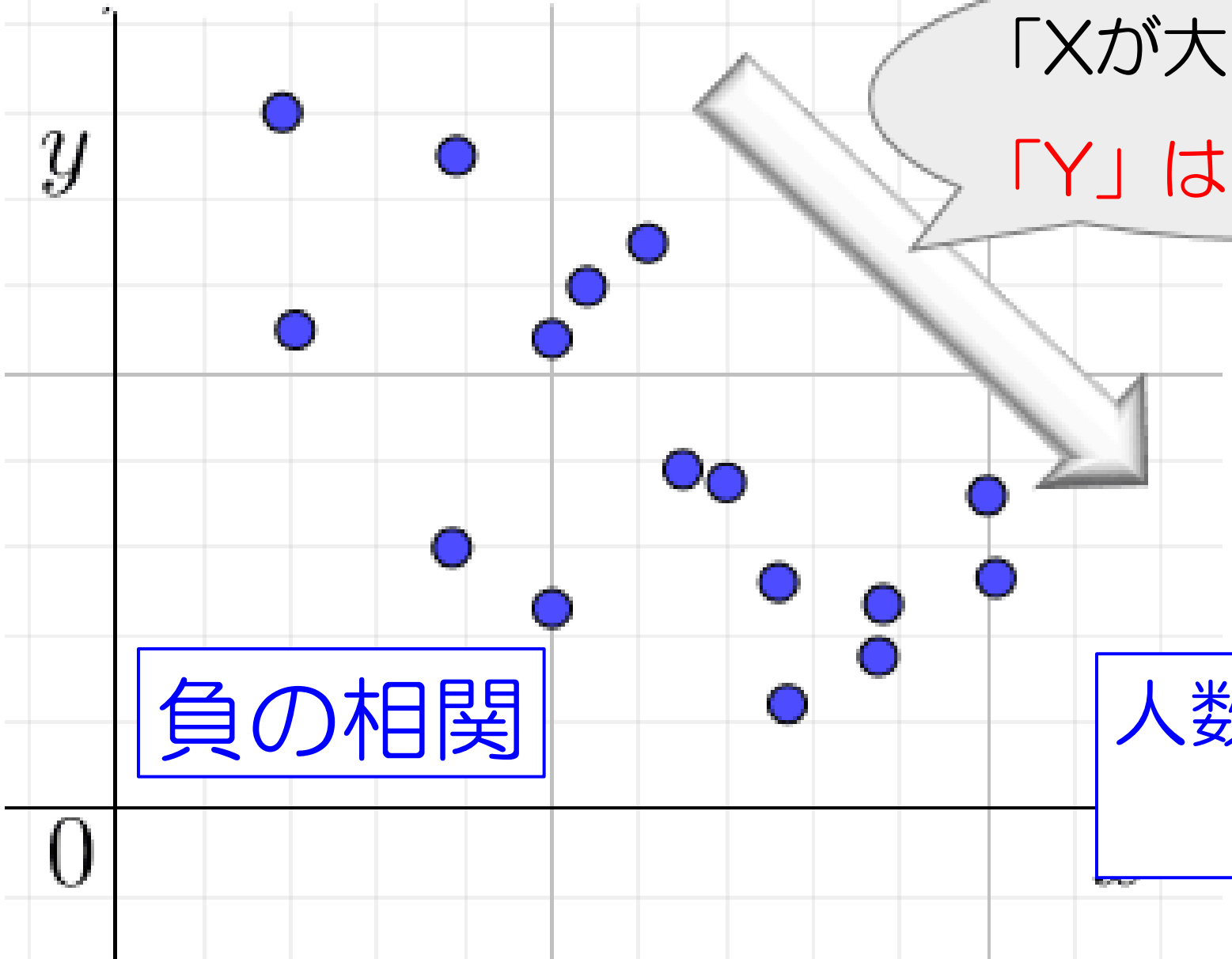
身長が大きくなると
体重も増える

正の相関



「Xが大きくなると
「Y」も大きくなる

「相関係数：−0.74」



「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

「相関係数の求め方」

$$r = \frac{Sxy}{Sx \times Sy} = \frac{\frac{1}{n}\sum_{i=0}^{n}(xi - x)(yi - y)}{\sqrt{\frac{1}{n}\sum_{i=0}^{n}(xi - x)^2} \times \sqrt{\frac{1}{n}\sum_{i=0}^{n}(yi - y)^2}}$$

r ：xとyの相関関数

Sxy ：xとyの共分散

Sx ： xの標準偏差

Sy ： yの標準偏差

めんどくさい！！

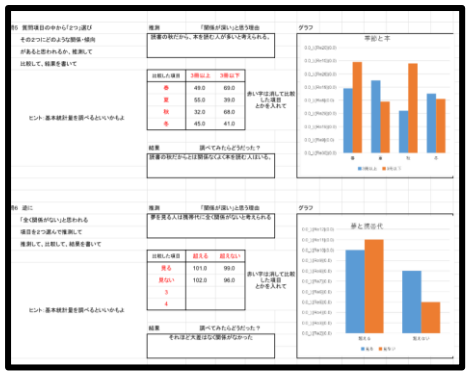
提出１

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！



問６　逆に
「全く関係がない」と思われる
項目を２つ選らんで推測して

- 「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！

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！

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