データの分析【散布図・相関係数】 p.193～201

青チャートⅠ：例題187～189

【態度目標】しゃべる、質問する、説明する、動く、協力する、貢献する

【内容目標】散らばりの度合いを表す値を求められるようになろう

□２つの変量の間の関係

気温と商品需要の関係，国語と英語の成績の関係など，2 つの変量の間の関係について調べたいことがある。ここで学ぶ相関関係とは，量的データをとる 2 つの変量の間の関係であり，散布図や相関係数によって調べることができる。また，質的データをとる 2 つの変量の間の関係についても調べよう。

□散布図

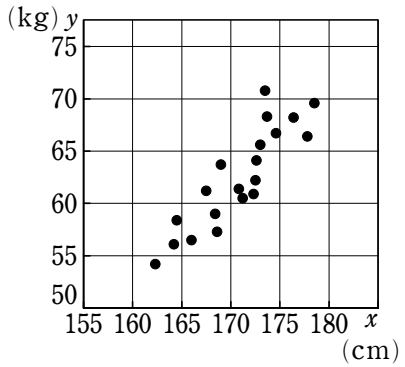
	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
x	168.4	164.5	171.2	173.0	162.3	170.8	172.5	164.2	169.0	168.6
y	59.0	58.4	60.5	65.6	54.2	61.4	62.2	56.1	63.7	57.3

	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱	⑲	⑳
x	172.6	166.0	173.7	176.4	178.5	167.5	177.8	174.6	172.3	173.5
y	64.1	56.5	68.3	68.2	69.6	61.2	66.4	66.7	60.9	70.8

上の表は，ある高校の 1 年生男子 20 人について，身長を x (cm)，体重を y(kg)として，x，y を調べた結果である。例えば，① の人は，身長 168.4 cm，体重 59.0 kg である。

x と y の間の関係を見やすくするために，右の図のように，x，y の値の組を座標とする点を平面上にとる。この図から，x が増えると y も増える傾向があることがわかる。

右のような図を **散布図** という。



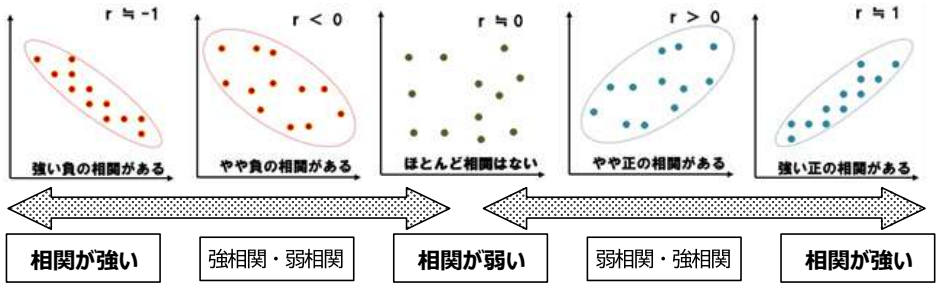
□相関関係

2 つの変量のデータにおいて，一方が増えると他方も増える傾向が認められるとき，2 つの変量の間に **正の相関関係** があるという。逆に，一方が増えると他方が減る傾向が認められるとき，2 つの変量の間に **負の相関関係** があるという。どちらの傾向も認められないときは，**相関関係がない** という。

有相関・無相関

補足 正の相関がある，負の相関がある，相関がない，ということもある。

2 つの変量の間に正の相関関係があるとき，散布図の点は全体に右上がりに分布し，負の相関関係があるときは，散布図の点は全体に右下がりに分布する。 2 つの変量の間に正の相関あるいは負の相関があるとき，散布図における点の分布が 1 つの直線に接近しているほど相関が強いといい，散らばっているほど相関が弱いという。



データの分析【散布図・相関係数】 p.193～201

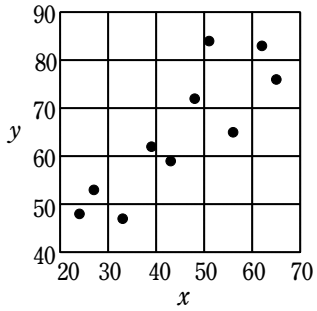
青チャートⅠ：例題187～189

例）

次のような 2 つの変量 x，y についてのデータがある。これらについて，散布図をかき，x と y の間に相関関係があるかどうかを調べよ。また，相関関係がある場合には，正・負のどちらであるかをいえ。

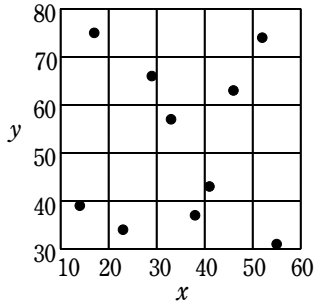
(1)

x	43	39	51	27	65	24	62	33	56	48
y	59	62	84	53	76	48	83	47	65	72



(2)

x	38	23	46	14	52	17	55	29	41	33
y	37	34	63	39	74	75	31	66	43	57



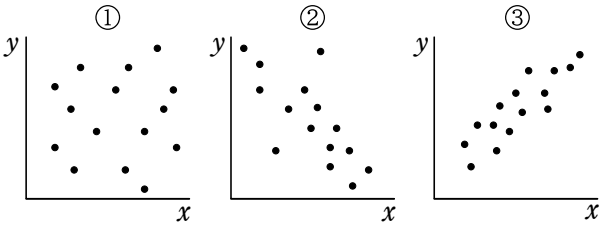
(1) 正の相関関係がある　　(2) 相関関係がない

例）

右の ①，②，③ は，ある 2 つの変量 x，y のデータについての散布図である。データ ①，②，③ の x と y の相関係数は，

0.87，0.04，−0.71

のいずれかである。各データの相関係数を答えよ。



0.87 ⇒ 正の相関、　0.04 ⇒ ほとんど相関がない
−0.71 ⇒ 負の相関　　と読み取る

解答　① 0.04　② −0.71　③ 0.87

データの分析【散布図・相関係数】 p.193～201

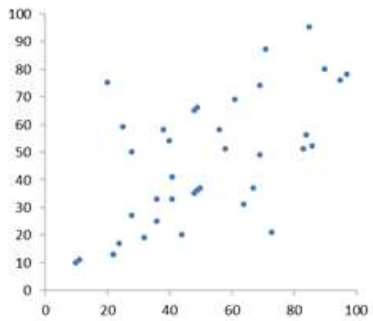
青チャートⅠ：例題187～189

□相関係数

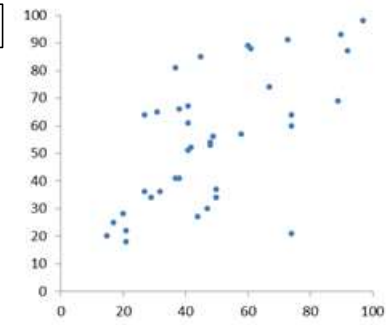
下の図は、ある２クラスの国語と英語のテストの散布図である。どの程度強い相関関係であるかを他の人に伝える際にどうすれば的確に伝えられるだろうか？

また微妙な散らばり具合の散布図の相関の強さを比較するとき、どうすれば判断することができるだろうか。

A組



B組



２つの変量からなるデータが与えられたとき、データの値から相関関係を調べる方法として、どの程度直線的であるかを数値で表すことで比較することができる。そこで相関関係の目安となる数値を考えよう。

2 つの変量 x，y のデータが，n 個の x，y の値の組として，次のように与えられているとする。

$$(x_1,\ y_1),\ (x_2,\ y_2),\ \cdots\cdots,\ (x_n,\ y_n)$$

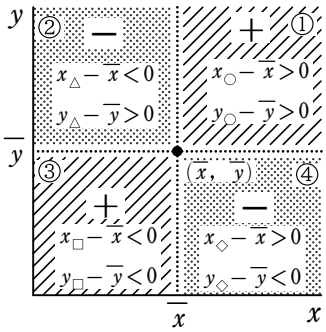
以下では，x_1，x_2，……，x_n と y_1，y_2，……，y_n の平均値をそれぞれ $\overline{x}$，$\overline{y}$，標準偏差をそれぞれ s_x，s_y とする。ここで，**x の偏差と y の偏差の積 $(x_k-\overline{x})(y_k-\overline{y})$ の平均値**

$$\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots\cdots+(x_n-\overline{x})(y_n-\overline{y})\}\quad\cdots\cdots①$$

を考える。① を x と y の **共分散** といい，s_{xy} で表す。

x_1，x_2，……，x_n の平均値を $\overline{x}$，y_1，y_2，……，y_n の平均値を $\overline{y}$ とし，$\overline{x}$，$\overline{y}$ を境界として，データの散布図を右の図のように ①，②，③，④ の領域に分ける。

$(x_k-\overline{x})(y_k-\overline{y})$ の符号



このとき，データの散布図について，次のことが考えられる。

点の多くが ① と ③ にあるとき，x と y の間に正の相関がある。

点の多くが ② と ④ にあるとき，x と y の間に負の相関がある。

ここで，たとえば，点 $(x_1,\ y_1)$ について考えてみると，

$(x_1,\ y_1)$ が ① または ③ にあるとき $(x_1-\overline{x})(y_1-\overline{y})>0$，

$(x_1,\ y_1)$ が ② または ④ にあるとき $(x_1-\overline{x})(y_1-\overline{y})<0$

である。よって，共分散の正負は，相関関係の正負の目安になると考えられる。

①～④に均等にあると
0に近くなる
よって相関関係が無い

⟹ x と y の間の相関を調べるのに，
x の偏差と y の偏差の積 （**共分散**） を用いると相関関係を調べられる

データの分析【散布図・相関係数】 p.193～201

青チャートⅠ：例題187～189

相関の強弱をみるために，共分散 s_{xy} を，x の標準偏差 s_x と y の標準偏差 s_y の積 $s_x s_y$ で割った値を考える。この値を x と y の **相関係数** といい，r で表す。

r は相関係数を母集団で計算したときのギリシャ文字 ρ(ロー)の英語表記 rho の頭文字

相関係数 r （correlation coefficient）

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

$\dfrac{(\boldsymbol{x}\ \boldsymbol{と}\ \boldsymbol{y}\ \boldsymbol{の共分散})}{(\boldsymbol{x}\ \boldsymbol{の標準偏差})\times(\boldsymbol{y}\ \boldsymbol{の標準偏差})}$ 【計算方法１】

$$=\frac{\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+\cdots\cdots+(x_n-\overline{x})(y_n-\overline{y})\}}{\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2+\cdots\cdots+(x_n-\overline{x})^2\}}\sqrt{\frac{1}{n}\{(y_1-\overline{y})^2+\cdots\cdots+(y_n-\overline{y})^2\}}}$$

$$=\frac{(x_1-\overline{x})(y_1-\overline{y})+\cdots\cdots+(x_n-\overline{x})(y_n-\overline{y})}{\sqrt{\{(x_1-\overline{x})^2+\cdots\cdots+(x_n-\overline{x})^2\}\{(y_1-\overline{y})^2+\cdots\cdots+(y_n-\overline{y})^2\}}}$$

$\dfrac{(\boldsymbol{x}\ \boldsymbol{の偏差と}\ \boldsymbol{y}\ \boldsymbol{の偏差の積の総和})}{\sqrt{(\boldsymbol{x}\ \boldsymbol{の偏差の}\ \boldsymbol{2}\ \boldsymbol{乗の総和})\times(\boldsymbol{y}\ \boldsymbol{の偏差の}\ \boldsymbol{2}\boldsymbol{乗}\ \boldsymbol{の総和})}}$ 【計算方法２】

相関係数 r については，次の性質がある。

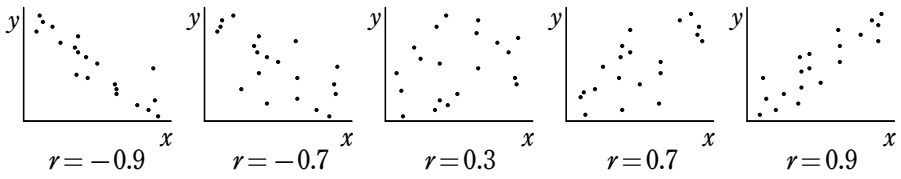
[1] $-1\leqq r\leqq 1$ ← 証明については後述

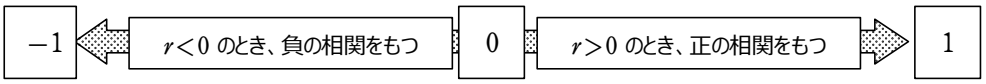
[2] $r=1$ のとき，散布図の点は右上がりの直線上に分布する。

[3] $r=-1$ のとき，散布図の点は右下がりの直線上に分布する。

[4] r の値が 0 に近いとき，直線的な相関関係はない。

また，r の値が 1 に近いことは正の相関関係が強いことの目安であり，r の値が -1 に近いことは負の相関関係が強いことの目安である。相関係数は x と y の直線的な相関関係を考察するための目安になる。





【参考】

相関係数	相関関係
0	相関がない
0.0～±0.2	ほとんど相関がない
±0.2～±0.4	やや相関がある(低い相関)
±0.4～±0.7	相関がある
±0.7～±0.9	強い相関がある(高い相関)
±0.9～±1.0	きわめて強い相関がある
±1.0	完全な相関

ちなみに前述のＡ組、Ｂ組の散布図の相関係数は

Ａ組：0.65 と Ｂ組：0.62 となるので、

微妙ながらＡ組の方が正の相関が強い

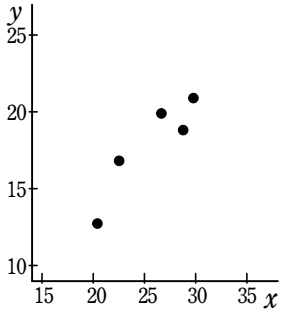
と言うことがわかる。

データの分析【散布図・相関係数】 p.193～201

青チャートⅠ：例題187～189

例９） 次の表は，同じ種類の５本の木について，根もとの太さ x (cm) と高さ y (m) を測定した結果である。x と y の相関係数 r を求めよう。

	①	②	③	④	⑤
x	21	27	29	23	30
y	13	20	19	17	21



散布図は右の図のようになる。

x，y のデータの平均値は

$\overline{x}=\frac{1}{5}\times 130=26$，$\overline{y}=\frac{1}{5}\times 90=18$

	x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})(y-\overline{y})$	$(x-\overline{x})^2$	$(y-\overline{y})^2$
①	21	13	−5	−5	25	25	25
②	27	20	1	2	2	1	4
③	29	19	3	1	3	9	1
④	23	17	−3	−1	3	9	1
⑤	30	21	4	3	12	16	9
計	130	90			(45)	△60	□40
平均	26	18			9	12	8

$$\frac{(\boldsymbol{x}\text{ の偏差と }\boldsymbol{y}\text{ の偏差の積の総和})}{\sqrt{(\boldsymbol{x}\text{ の偏差の 2 乗の総和})\times(\boldsymbol{y}\text{ の偏差の 2乗 の総和})}}=\frac{○}{\sqrt{△\times□}}$$ 【計算方法２】を利用

上の表から，相関係数 r は $r=\frac{45}{\sqrt{60\times 40}}\fallingdotseq 0.92$

相関係数 r が正で 1 に近いから，x と y には強い正の相関があると考えられる。 終

【参考】相関表

散布図にすると重なる点がある場合、正しく表すことができない。

またデータの数が多すぎると点だらけになってしまう。

そのようなときは相関表を用いると良い。

	0 ～ 20	20 ～ 40	40 ～ 60	60 ～ 80	80 ～ 100	計
80以上100未満			2	3	6	11
60 ～ 80		1	3	14	4	22
40 ～ 60	2		6	12	2	22
20 ～ 40		4	2			6
0 ～ 20	5	1		1		7
計	7	6	13	30	12	68

データの分析【散布図・相関係数】 p.193～201

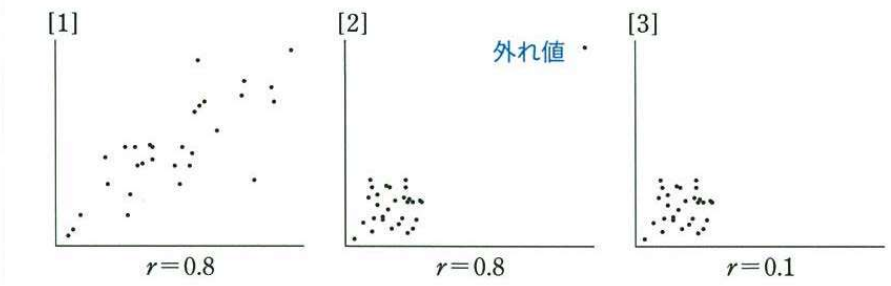
青チャートⅠ：例題187～189

相関係数を利用してデータを分析するとき，その数値だけで判断しないように注意する。

□ 相関係数の値と散布図

相関係数は，外れ値の影響を受けやすい値である。

下の散布図 [1]，[2] で表されたデータは，いずれも相関係数が 0.8 である。しかし，[2] のデータから外れ値を 1 つだけ除いたデータ [3] の相関係数は 0.1 である。2 つの変量の間の相関係数を調べるとき，相関係数の値だけでは，分布の特徴を正しくとらえられない場合もある。



補足　$-1 \leqq r \leqq 1$ であることの証明

$a_k = x_k - \overline{x}$，$b_k = y_k - \overline{y}$ $(k=1, 2, \cdots\cdots, n)$ と任意の実数 t に対して，

$\sum_{k=1}^{n} (a_k t - b_k)^2 \geqq 0$ が成り立つから

$$\left(\sum_{k=1}^{n} a_k{}^2\right)t^2 - 2\left(\sum_{k=1}^{n} a_k b_k\right)t + \left(\sum_{k=1}^{n} b_k{}^2\right) \geqq 0$$

が成り立つ。よって，その判別式 D について

$$\frac{D}{4} = \left(\sum_{k=1}^{n} a_k b_k\right)^2 - \left(\sum_{k=1}^{n} a_k{}^2\right)\left(\sum_{k=1}^{n} b_k{}^2\right) \leqq 0$$

が成り立つ。したがって

$$r^2 = \frac{\left(\sum_{k=1}^{n} a_k b_k\right)^2}{\left(\sum_{k=1}^{n} a_k{}^2\right)\left(\sum_{k=1}^{n} b_k{}^2\right)} \leqq 1$$

これにより，$|r| \leqq 1$ が示せる。

コーシー・シュワルツの不等式

任意の実数 a_i，b_i に対して

$(a_1{}^2 + a_2{}^2)(b_1{}^2 + b_2{}^2) \geqq (a_1 b_1 + a_2 b_2)^2$

$(a_1{}^2 + a_2{}^2 + a_3{}^2)(b_1{}^2 + b_2{}^2 + b_3{}^2) \geqq (a_1 b_1 + a_2 b_2 + a_3 b_3)^2$

という不等式が成立する。より一般に，

任意の正の整数 n に対して

$$\left(\sum_{i=1}^{n} a_i{}^2\right)\left(\sum_{i=1}^{n} b_i{}^2\right) \geqq \left(\sum_{i=1}^{n} a_i b_i\right)^2$$

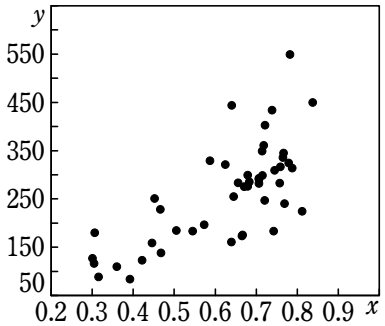
という不等式が成立する

データの分析【散布図・相関係数】 p.193～201

青チャートⅠ：例題187～189

□相関関係と因果関係

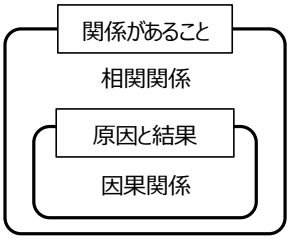
47 都道府県において，2017 年度の森林面積の総面積に対する割合 x と，人口 100万人あたりの郵便局の数 y のデータを調べ，散布図にしたところ，右の図のようになった。相関係数は 0.72 である。



（「政府統計の総合窓口（e−Stat）」，「林野庁ホームページ」より作成）

x と y の間には正の相関関係が認められる。しかし，森林面積の割合が多いことが原因で郵便局が増えるとか，逆に，郵便局が多いことが原因で森林面積の割合が大きくなるということは断定できないであろう。つまり，一方が原因で他方が起こる **因果関係** があるとは断定できない。

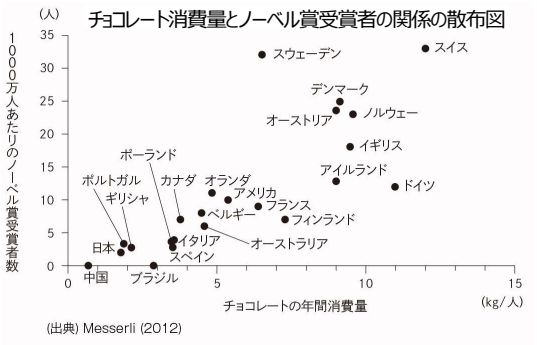
一般に，2 つの変量の間に相関関係があるからといって，必ずしも因果関係があるとはいえない。

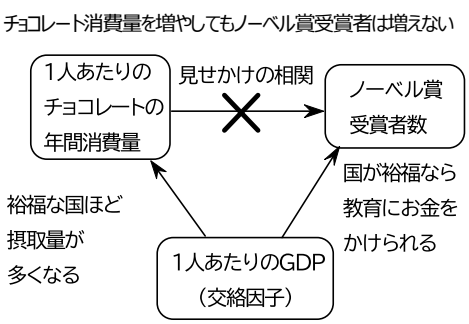


疑似相関とは？

因果関係がないのに、見えない要因によってあたかも因果関係があるように見える現象のこと。統計学でよく使われる言葉で、「見せかけの相関」「見かけ上の相関」とも言う。

一般的に有名な疑似相関の例としては、「チョコレートの消費量とノーベル賞の受賞者数」、「アイスクリームの売上と水難事故の数」、「年賀状を出す枚数と収入の高さ」等がある。





チョコレートは生きていくのになくてもよいいわゆる贅沢品であるので、裕福な国ほど摂取量が多くなるのは当然である。また国が裕福になれば、教育にもお金をかけられるようになるので、ノーベル賞受賞者を輩出できる可能性は上がると考えられる。

このような「第3の変数」のことを「**交絡因子（こうらくいんし）**」と呼ぶ。この交絡因子があると、相関関係にすぎないものがまるで因果関係のように見えてしまう。

データの分析【散布図・相関係数】 p.193～201

青チャートⅠ：例題187～189

□質的データをとる 2 つの変量の間の関係

質的データをとる 2 つの変量の間の関係を調べることについて考えよう。

例１０）

合否が判定されるある試験において，受験者 100 人全員を対象に，教材 A を使用して学習したか調べたところ，その人数は表 1 のようになった。例１０ の表 1 のような表を **分割表** という。クロス集計表ともいう。

表 1　**分割表**

	合	否	計
A の使用：有	9	5	14
A の使用：無	42	44	86
計	51	49	100

$9 \div 14 = 0.6428\cdots$　$5 \div 14 = 0.3571\cdots$

$42 \div 86 = 0.4883\cdots$　$44 \div 86 = 0.5116\cdots$

表 1 において，教材 A を使用した者，使用していない者のそれぞれにおいて，合格者，不合格者が占める割合を計算すると，表 2 のようになる。

表 2

	合	否
A：有	64%	36%
A：無	49%	51%

表 2 だけを見ると，教材 A の使用が合否に影響を及ぼしているようにみえる。ここでさらに，教材 B を使用して学習したかも調べてみたところ，その人数は表 3 のようになった。

表 3

		合	否
A：有	B：有	6	1
A：有	B：無	3	4
A：無	B：有	31	9
A：無	B：無	11	35

練習１５）　表 3 をもとに，次の問いに答えよ。

(1)　表 4 の空らんに適切な数を入れよ。

表 4

	合	否	計
B：有	37	10	47
B：無	14	39	53
計	51	49	100

$37 \div 47 = 0.7872\cdots$　$10 \div 47 = 0.2127\cdots$

$14 \div 53 = 0.2641\cdots$　$39 \div 53 = 0.7358\cdots$

(2)　教材 B を使用した者，使用していない者のそれぞれにおいて，合格者，不合格者の占める割合を計算して表 2 のようにまとめよ。

	合	否
B：有	79%	21%
B：無	26%	74%

深める　上で得られたデータから，教材 A，B のどちらの方が，この試験の合否により影響を及ぼしていると予想できるだろうか。

	合	否
A：有	64%	36%
A：無	49%	51%

	合	否
B：有	79%	21%
B：無	26%	74%

		合	否
A：有	B：有	6	1
A：有	B：無	3	4
A：無	B：有	31	9
A：無	B：無	11	35

解答

使用している受検者の合格する割合が多いことから、Bの方がより影響を及ぼしていると予想できる。（Aは無くても受かっている受検者も多く、そもそも使用している人数が少なく判断しづらい）

データの分析【散布図・相関係数】 p.193～201

青チャートⅠ：例題187～189

参考 オッズ比を用いて，事象の起こりやすさを比較することもできる。

ある事柄が起こる割合 (p) を，その事象が起こらない割合 $(1-p)$ で割ったものをオッズ比という。

例１０では表１において　Aの使用有りのオッズは $\dfrac{\frac{9}{14}}{\frac{5}{14}}=\dfrac{9}{5}$ ，Aの使用無しのオッズは $\dfrac{\frac{42}{86}}{\frac{44}{86}}=\dfrac{42}{44}$ で

あり，それらの比の値 $\dfrac{\frac{9}{5}}{\frac{42}{44}}(\fallingdotseq 1.89)$ がオッズ比である。Bも同様にオッズ比を求めると $\dfrac{\frac{37}{10}}{\frac{14}{39}}(\fallingdotseq 10.31)$ である。

オッズ比の値はBの方が大きく　Bの方がこの試験に合格することにより影響を与えていると考察できる。

研究 最小２乗法

実際のデータはある程度散らばっているため、どの部分に直線を引くのが妥当かが問題となる。その解決方法の一つとして「**最小２乗法**」がある。これはガウスによって体系的に理論化され、未知の真の値を観測によって推定する方法として有効であるとされている。最小２乗法は，人工知能（AI）の分野でも用いられる。

データ d_1，d_2，d_3，d_4，d_5 が与えられたとき，データとの差の２乗の和

$(d_1-x)^2+(d_2-x)^2+(d_3-x)^2+(d_4-x)^2+(d_5-x)^2$ …… ①

を最小にする x の値について考えてみよう。① を x の関数とみて $f(x)$ とおくと

$$f(x)=(d_1-x)^2+(d_2-x)^2+(d_3-x)^2+(d_4-x)^2+(d_5-x)^2$$
$$=5x^2-2(d_1+d_2+d_3+d_4+d_5)x+{d_1}^2+{d_2}^2+{d_3}^2+{d_4}^2+{d_5}^2$$

ここで，　$a=\dfrac{d_1+d_2+d_3+d_4+d_5}{5}$　(データの平均値)

$b=\dfrac{{d_1}^2+{d_2}^2+{d_3}^2+{d_4}^2+{d_5}^2}{5}$　(データの２乗の平均値)

とおくと　$f(x)=5x^2-2\cdot 5a\cdot x+5b=5(x-a)^2+5(b-a^2)$

よって，$f(x)$ は x の２次関数で，$x=a$ で最小値 $5(b-a^2)$ をとる。

ここで，最小値を与える x の値 a はデータの平均値であり，最小値をデータの大きさ５で割って得られる $b-a^2$ はデータの分散である。

データとの差の２乗の和を最小にするものを求めることは，観測値から未知の真の値を推定する方法として広く用いられている。これを **最小２乗法** という。

また、このように、最小２乗法によってデータに最もよくあてはまる直線を求めることを、

線形回帰（linear regression）といい，得られた直線を**回帰直線**という。

最小２乗法で求められた直線の方程式の係数は，平均値，標準偏差，相関係数で表される。つまり，２つの変量 x，y に対して最小２乗法による回帰直線の方程式は

$$y-\overline{y}=r\cdot\frac{s_y}{s_x}(x-\overline{x}) \quad \text{すなわち} \quad \frac{y-\overline{y}}{s_y}=r\cdot\frac{x-\overline{x}}{s_x}$$

青チャートp.319

$\overline{x}$，$\overline{y}$：それぞれの x，y の平均値　s_x，s_y：それぞれの x，y の標準偏差　r：x と y の相関係数

となり，回帰直線は点 $(\overline{x}, \overline{y})$ を通ることがわかる。

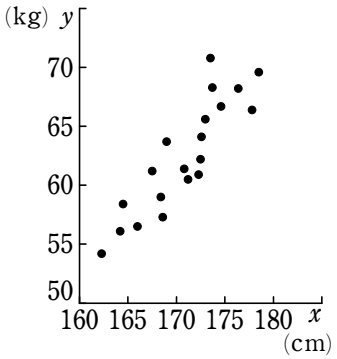
コラム

回帰分析

青チャートp.319

ある高校の１年生男子の身長 x と体重 y の散布図について，これらの点は，ある直線の近くに並んでいるようにも見える。

そこで，このデータの傾向を最もよく表す１次関数を見つけることを考えよう。



散布図において，点の配列に「できるだけ合うように引いた直線」を**回帰直線**という。そこで，この回帰直線をこの散布図の中に引くことを考える。

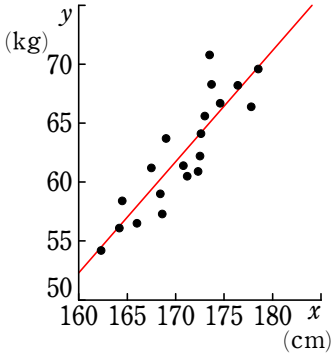
直線を引く基本的な方法は回帰分析と呼ばれるもので，コンピュータなどの情報機器を利用してかくこともできる。

> 実際には手計算で回帰直線を引くことは少なく、
> 表計算ソフトなどを利用することがほとんどである。

上の x と y のデータでは，次の１次関数が得られる。

$$y=0.943x-98.603$$

実際に直線を引くと右の図のようになる。



回帰分析は，自然科学のデータ分析で必須であるだけでなく，経済学や社会学などの社会科学を学ぶ上でも重要な手法である。

> 回帰直線を引くことで、１つの変量のデータから
> もう一方の変量のデータの値を予測することが可能である。

データの分析【散布図・相関係数】 p.193～201

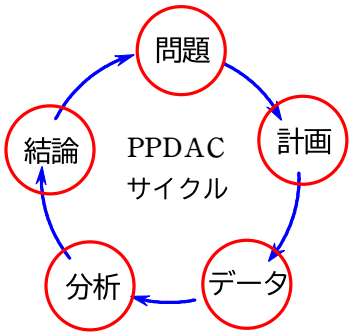
青チャートⅠ：例題187～189

研究　統計的探究プロセス

実社会では，さまざまな社会的問題に応じて，統計的手法を用いた問題解決が行われている。そのときには，

「問題 → 計画 → データ → 分析 → 結論」

の 5 段階からなる **統計的探究プロセス** を意識することが大事である。



問 題（Problem） … 解決すべき事柄を把握し，統計で扱える問題を設定する。

計 画（Plan） … 設定した問題に対して，集めるべきデータと集め方を考える。

データ（Data） … 計画にしたがってデータを集め，表などに整理する。

分 析（Analysis） … 目的やデータの種類に応じてグラフにまとめたり，データに関する数値を求めたりして，特徴や傾向を把握する。

結 論（Conclusion） … 見いだした特徴や傾向から結論をまとめて表現したり，さらなる課題や改善点を見いだしたりする。

また，実社会でのデータは，一般に非常に大量であり，手計算では処理しきれないことがほとんどである。そのような大量のデータを扱う際には，コンピュータなどの情報機器を用いて，グラフをかいたり，さまざまな計算を行うとよい。

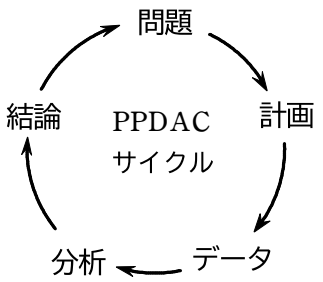
データの分析【散布図・相関係数】 p.193～201

青チャートⅠ：例題187～189

改良したボールペンＢを売り出すための方策を統計的プロセスを適用して考えてみよう。

問　題（Problem） 解答

改良したボールペンＢを積極的に売り出すために，
ボールペンＢが既存のボールペンＡより書きやすく
改良されているかを，データから客観的に示したい。
すなわち，主張「①ＡよりＢの方が書きやすい」を立てて，
①が正しいことを統計的に示したい。



計　画（Plan）

すべての消費者集めるのは困難のため，街中で無作為に人を集めてＡとＢを使ってもらい，どちらが書きやすいかデータを集めることにした。

なお，先入観をもたせないため，どちらが改良したボールペンかは伝えないようにすることにした。

データ（Data）

実際に街中でデータを集め，30 人にアンケートを取りデータを集めた。その結果，30 人中 21 人がＢの方が書きやすいと答えた。

分　析（Analysis）

データの結果が統計的にどのような意味をもつか検証したい。ここでは主張①に反する仮定「②　Ａ，Ｂのどちらの回答も全くの偶然で起こる　」を考え，②のもと 30 人中 21 人以上がＢと回答する確率を見積もることにした。（例えば，コンピュータのシミュレーションで）その確率が 0.02 程度であることを特定できた。

結　論（Conclusion）

仮定②のもとで 30 人中 21 人以上Ｂと回答する確率が 0.02 程度と非常に低いことから，仮定②が　正しくなかった　と考えられ，主張①が　正しい　と判断してもよさそうである。すなわち，ボールペン　Ｂ　の方が書きやすいと主張できる。

よかった点

調査のときに　どちらのボールペンが改良したものかを伝えない　ことで，余計な先入観を入れずデータが集められたと考えられ，それからもこの結論は妥当だろうと考えてもよいだろう。

課題や改善点

○確率 0.02 で仮定②を正しくないと判断したが，この確率の基準を事前に設定する方がより客観的になるのではないか。

○今回の分析法で，30 人のデータから②が正しいとどのくらいの確率で判断できるだろうか。

○標本数として 30 人は妥当だろうか。

○もっと多くの標本を集めた場合は，統計的により高い精度が保証できるだろうか。

データの分析【散布図・相関係数】 p.193～201

青チャートⅠ：例題187～189

参考 相関係数の別な計算の仕方

注意 以下の話は数列で勉強するシグマを使用していますので，未履修の場合は飛ばしてください。

数列 $\{a_n\}$ について，初項から第 n 項までの和を，記号 $\sum$ を用いて $\sum_{k=1}^{n} a_k$ と書く。

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + a_3 + \cdots\cdots + a_n$$

$$(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots\cdots+(x_n-\overline{x})^2=\sum_{k=1}^{n}(x_k-\overline{x})^2$$

$$\sum_{k=1}^{n}(x_k-\overline{x})^2=\sum_{k=1}^{n}x_k{}^2-n(\overline{x})^2$$

$$(y_1-\overline{y})^2+(y_2-\overline{y})^2+\cdots\cdots+(y_n-\overline{y})^2=\sum_{k=1}^{n}(y_k-\overline{y})^2$$

$$\sum_{k=1}^{n}(y_k-\overline{y})^2=\sum_{k=1}^{n}y_k{}^2-n(\overline{y})^2$$

$$\begin{aligned}
&\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})\\
&=\sum_{k=1}^{n}(x_ky_k-\overline{x}y_k-\overline{y}x_k+\overline{x}\cdot\overline{y})\\
&=\sum_{k=1}^{n}x_ky_k-\overline{x}\sum_{k=1}^{n}y_k-\overline{y}\sum_{k=1}^{n}x_k+n\overline{x}\cdot\overline{y}\\
&=\sum_{k=1}^{n}x_ky_k-\overline{x}\cdot n\cdot\frac{1}{n}\sum_{k=1}^{n}y_k-\overline{y}\cdot n\cdot\frac{1}{n}\sum_{k=1}^{n}x_k+n\overline{x}\cdot\overline{y}\\
&=\sum_{k=1}^{n}x_ky_k-n\overline{x}\cdot\overline{y}-n\overline{y}\cdot\overline{x}+n\overline{x}\cdot\overline{y}\\
&=\sum_{k=1}^{n}x_ky_k-n\overline{x}\cdot\overline{y}
\end{aligned}$$

よって次の公式が成り立つ。

$$r=\frac{(x_1-\overline{x})(y_1-\overline{y})+\cdots\cdots+(x_n-\overline{x})(y_n-\overline{y})}{\sqrt{\{(x_1-\overline{x})^2+\cdots\cdots+(x_n-\overline{x})^2\}\{(y_1-\overline{y})^2+\cdots\cdots+(y_n-\overline{y})^2\}}}=\frac{\sum_{k=1}^{n}(x_ky_k-n\overline{x}\cdot\overline{y})}{\sqrt{\left\{\sum_{k=1}^{n}(x_k{}^2-n(\overline{x})^2)\right\}\left\{\sum_{k=1}^{n}(y_k{}^2-n(y)^2)\right\}}}$$