§２　２次元のデータ

2.1　相関図

相関図，相関関係とは？

○２つの変数の関係を座標平面上の点で表したものを**散布図**という。

○２つの変数の間に

一方が増えると他方も増える傾向があるとき，２つの変量の間に

正の相関関係があるといい，

一方が増えると他方は減る傾向があるとき，２つの変量の間に

負の相関関係があるという。

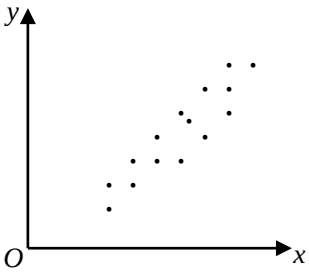
○正・負いずれの相関関係も見られないとき、

相関関係はないという。

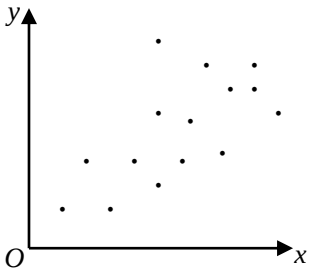
○２つの変数の間に相関関係があり，特に散布図の直線的傾向が強いとき，

相関関係が強いといい，直線的傾向が弱く散らばっているとき，

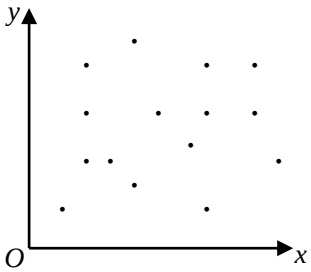
相関関係が弱いという。



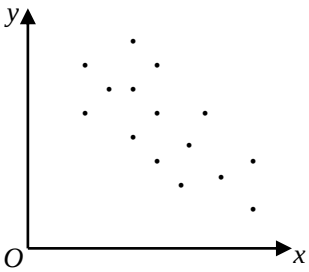
強い正の相関関係



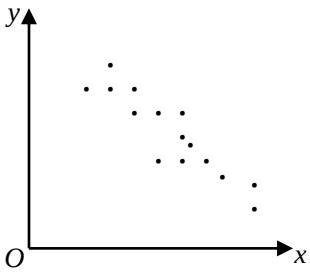
弱い正の相関関係



相関関係はない



弱い負の相関関係



強い負の相関関係

2.2　共分散と相関係数

共分散，相関係数とは？

○２つの変数 x，y が N 組のデータ

$(x_1,\ y_1),\ (x_2,\ y_2),\ (x_3,\ y_3),\ \cdots,\ (x_N,\ y_N)$

として，与えられているとき，次の式で表される変数を
２つの変数 x，y の**共分散**といい，記号 σ_{xy} で表す。

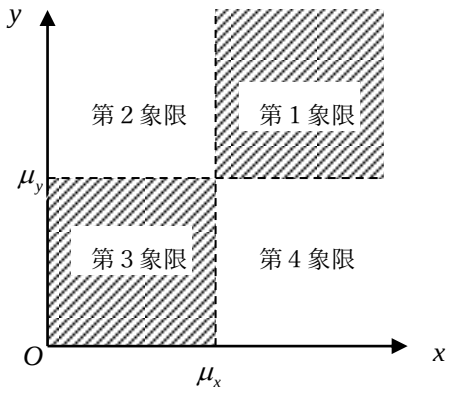
［共分散］　$\sigma_{xy}=\dfrac{1}{N}\sum_{k=1}^{N}(x_k-\mu_x)(y_k-\mu_y)$

$=\dfrac{(x_1-\mu_x)(y_1-\mu_y)+(x_2-\mu_x)(y_2-\mu_y)+\cdots+(x_N-\mu_x)(y_N-\mu_y)}{N}$

但し，μ_x，μ_y は各変数 x，y の平均値

【注意：共分散と散布図の関係】

各 $(x_k-\mu_x)(y_k-\mu_y)$ の値は，
第１象限と第３象限では正となり，
第２象限と第４象限では負になる。



共分散はこれらの和から
構成されていることから，
共分散の値が正のときは，正の相関関係があると考えられ，
共分散の値が負のときは，負の相関関係があると考えられる。

○２つの変数 x，y の共分散を，各変数 x，y の標準偏差 σ_x，σ_y の
積 $\sigma_x\sigma_y$ で割ったものを**相関係数** $r=\dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$ という。

○相関係数 r については，次のことが指摘されます。

・相関係数は，-1 から 1 までの値。（全範囲：$-1\leqq r\leqq 1$）
・相関係数が，1 に近いときは強い正の相関関係。（範囲：$0.8\leqq r\leqq 1$）
・相関係数が，-1 に近いときは強い負の相関関係。（範囲：$-1\leqq r\leqq -0.8$）
・相関係数が，0 に近いときは，相関関係はない。（範囲：$-0.2\leqq r\leqq 0.2$）

○共分散は，次の計算式でも求めることができる。

［共分散］　$\sigma_{xy} = E(xy) - E(x)E(y)$

但し　$E(x) = \mu_x = \dfrac{1}{N}\sum_{k=1}^{N} x_k = \dfrac{x_1 + x_2 + \cdots + x_N}{N}$

$E(y) = \mu_y = \dfrac{1}{N}\sum_{k=1}^{N} y_k = \dfrac{y_1 + y_2 + \cdots + y_N}{N}$

$E(xy) = \dfrac{1}{N}\sum_{k=1}^{N} x_k y_k = \dfrac{x_1 y_1 + x_2 y_2 + \cdots + x_N y_N}{N}$

⇩（※ y を x に書き換えると，次の式になる）

【復習：分散と２乗平均】

［分散］　$V(x) = \sigma_x^2 = E(x^2) - \{E(x)\}^2$

$E(x) = \dfrac{1}{N}\sum_{k=1}^{N} x_k = \dfrac{x_1 + x_2 + \cdots + x_N}{N}$

$E(x^2) = \dfrac{1}{N}\sum_{k=1}^{N} x_k^2 = \dfrac{x_1^2 + x_2^2 + \cdots + x_N^2}{N}$

[共分散の証明]

$$\sigma_{xy} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu_x)(y_k - \mu_y) = \frac{1}{N}\sum_{k=1}^{N}(x_k y_k - \mu_y x_k - \mu_x y_k + \mu_x \mu_y)$$

$$= \frac{1}{N}\sum_{k=1}^{N} x_k y_k - \mu_y \times \frac{1}{N}\sum_{k=1}^{N} x_k - \mu_x \times \frac{1}{N}\sum_{k=1}^{N} y_k + \mu_x \mu_y \times \frac{1}{N}\sum_{k=1}^{N} 1$$

$$= E(xy) - \mu_y \times \mu_x - \mu_x \times \mu_y + \mu_x \mu_y \times \frac{N}{N}$$

$$= E(xy) - \mu_x \mu_y = E(xy) - E(x)E(y)$$

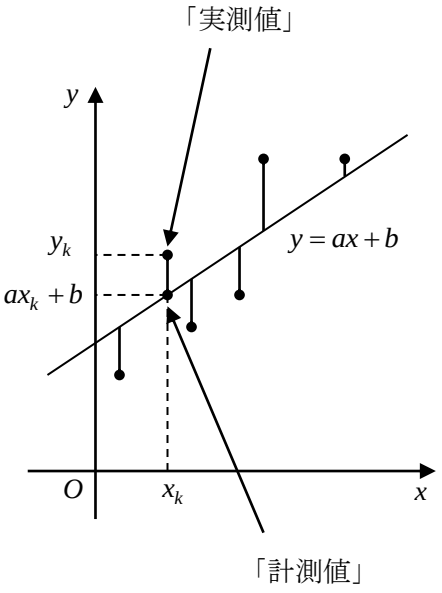
2.3　回帰直線

最小２乗法，回帰直線とは？

○２つの変数 x，y の N 組のデータ

$(x_1,\ y_1),\ (x_2,\ y_2),\ \cdots,\ (x_N,\ y_N)$

において，
各データと直線 $y=ax+b$ との差が
最小になるように a，b を定める
方法の一つとして**最小２乗法**がある。



○これは，変数 y の「実測値」 y_k と
変数 x から求められる変数 y の
「予測値」 ax_k+b との差

$$\varepsilon_k = y_k - (ax_k + b)$$

を２乗したものの和

$$\varepsilon = \sum_{k=1}^{N} \varepsilon_k^2 = \sum_{k=1}^{N} \{y_k - (ax_k + b)\}^2$$

を最小にするものです。[ギリシア文字 ε の読み：イプシロン]

○最小２乗法により定まる a，b を係数とする直線を，**回帰直線**といい，
次の式で，与えられることが知られている。
[※証明には偏微分の極値問題を解く必要があるので割愛します。]

［回帰直線］　$y-\mu_y=\dfrac{\sigma_{xy}}{\sigma_x^2}(x-\mu_x)$

【注意】回帰直線の傾き $m=\dfrac{\sigma_{xy}}{\boxed{\sigma_x^2}}$ と相関係数 $r=\dfrac{\sigma_{xy}}{\boxed{\sigma_x\sigma_y}}$ を混同しない！

※違う！！

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

【復習：直線の方程式】

点 $(a,\ b)$ を通る傾きが m の直線の方程式

［直線の方程式］　$y-b=m(x-a)$

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

例題　下の表は，10 人の学生の科目 A と科目 B の試験成績(10 点満点)である。

番号	1	2	3	4	5	6	7	8	9	10
A	7	8	4	9	6	5	8	7	8	8
B	8	10	5	8	10	7	9	9	8	6

科目 A のデータを変数 x，科目 B のデータを変数 y とするとき，
次の問いに答えよ。

(1) 科目 A の平均値 μ_x と分散 σ_x^2 を求めよ。

(2) 科目 B の平均値 μ_y と分散 σ_y^2 を求めよ。

(3) 科目 A と科目 B の共分散 σ_{xy} を求めよ。

(4) 科目 A と科目 B の相関係数 r を求めよ。

(5) 回帰直線を求めよ。

(6) 散布図を作成し，その中に回帰直線も描け。

［解答］

番号	x	y	$x-\mu_x$	$y-\mu_y$	$(x-\mu_x)^2$	$(y-\mu_y)^2$	$(x-\mu_x)(y-\mu_y)$
1	7	8	0	0	0	0	0
2	8	10	1	2	1	4	2
3	4	5	–3	–3	9	9	9
4	9	8	2	0	4	0	0
5	6	10	–1	2	1	4	–2
6	5	7	–2	–1	4	1	2
7	8	9	1	1	1	1	1
8	7	9	0	1	0	1	0
9	8	8	1	0	1	0	0
10	8	6	1	–2	1	4	–2
合計	70	80			22	24	10

(1) 科目 A :　平均値　$\mu_x = \dfrac{70}{10} = 7$　　分散　$\sigma_x^2 = \dfrac{22}{10} = 2.2$

(2) 科目 B :　平均値　$\mu_y = \dfrac{80}{10} = 8$　　分散　$\sigma_y^2 = \dfrac{24}{10} = 2.4$

(3) 共分散　$\sigma_{xy} = \dfrac{10}{10} = 1$

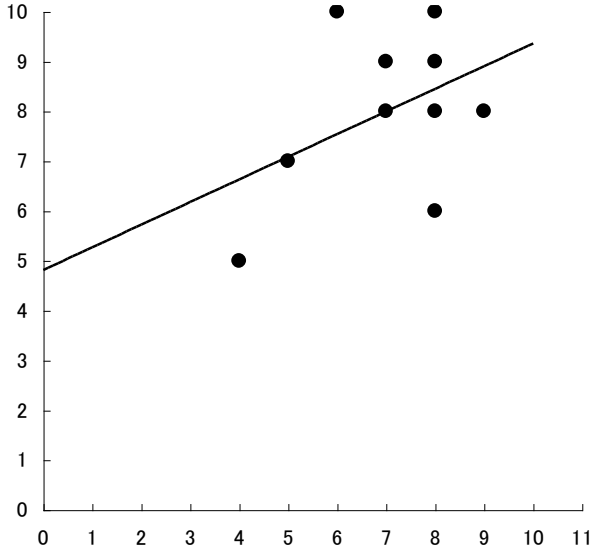
(4) 相関係数　$\sigma_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = \dfrac{1}{\sqrt{2.2}\sqrt{2.4}} \fallingdotseq \dfrac{1}{\sqrt{2.3}\sqrt{2.3}} = \dfrac{1}{2.3} = 0.43$

（※近似計算を行った）

(5) 回帰直線の傾き　$m = \dfrac{\sigma_{xy}}{\sigma_x^2} = \dfrac{1}{2.2}$

よって，求める回帰直線は　$y - 8 = \dfrac{x - 7}{2.2}$　（$\Rightarrow$　$y = 0.45x + 4.82$）

(6)



※データ数が少ない
こともあますが，
相関係数の値
[弱い正の相関関係]
を見ても
相関関係が「ない」
と否定もできないし，
相関関係が「ある」
と断言もできません。

問 3.9　下の 10 個の 2 次元データ

(3, 6)　(3, 7)　(4, 5)　(5, 5)　(6, 6)

(7, 4)　(8, 4)　(8, 4)　(8, 4)　(8, 5)

において，次の問いに答えよ。

(1) 平均値 μ_x と分散 σ_x^2 を求めよ。

(2) 平均値 μ_y と分散 σ_y^2 を求めよ。

(3) 共分散 σ_{xy} を求めよ。

(4) 相関係数 r を求めよ。

(5) 回帰直線を求めよ。

(6) 散布図を作成し，その中に回帰直線も描け。

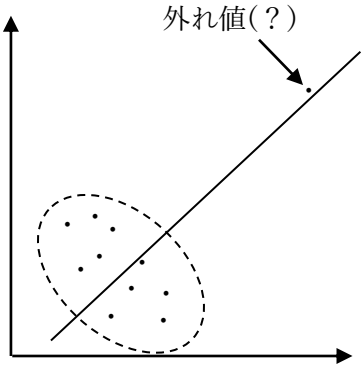
□□□□【参考】□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

２つの内容をお話します。
まずは**誤差**についてです。

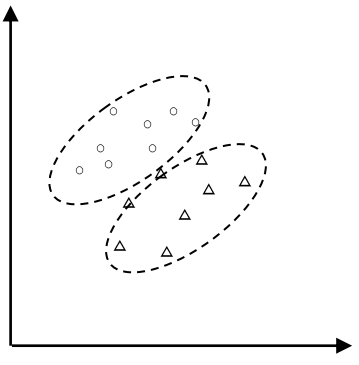
○測定や計算によって得られる値と，真値や正しい理論値との差を**誤差**と言います。
○実験など測定により生じる誤差[実測値と真値との差]を**測定誤差**,
データを用いて計算する際に生じる誤差[計算値と理論値との差]を**計算誤差**
と言います。
○統計では，**平均値**からどれくらいのばらつきがあるかを示す**標準偏差**
[分散の平方根]は，統計量の標準的な誤差を表しているとも言えます。
○また，データ数を多くすること，同じ実験や調査を繰り返し行うことなどで,
統計量が表す内容の信頼性や信憑性を高めることができます。

次は，相関図についてです。

○「外れ値(？)」には注意が必要！！
右の 10 個のデータの散布図において
全データを用いると
右上りの回帰直線がでます。
しかし，右上１個のデータを除くと
負の相関関係がありそうです。
右上１個のデータが外れ値かどうかは
場合によっては，同一条件で別集団での
再検討が必要かもしれません。



○層別を取り入れる工夫も必要
データ数を増やすために，何も考えず取り
入れると大変なことになることもあります。
右の 16 個のデータの散布図において
全データを用いると
相関関係はないよう見えます。
でも○[例：２年]，△[例：１年]と区別する
[層別化する]と正の強い相関関係が見えます。



□□