データ解析

http://coconut.sys.eng.shizuoka.ac.jp/data/

静岡大学工学部
安藤和敏

2005.10.12

多変量データ

個体名　変数名　変数名　…

社員 No	社交性	勤勉性	企画力	判断力	判断力
1	7	6	7	8	10
2	4	5	5	4	4
3	6	8	4	4	8
4	5	5	5	5	8
5	6	6	4	5	6
6	6	5	6	6	7
7	4	4	6	6	8

個体

多変量データ

個体	変数 x	変数 y	変数 z	変数 w
1	x_1	y_1	z_1	w_1
2	x_2	y_2	z_2	w_2
⋮	⋮	⋮	⋮	⋮
n	x_n	y_n	z_n	w_n

平均値 $\bar{x}$

個体名	変数 x
1	x_1
2	x_2
⋮	⋮
n	x_n

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

同じ平均値を持つ3つのデータ

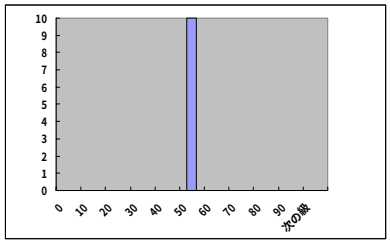
No	x
1	50
2	50
3	50
4	50
5	50
6	50
7	50
8	50
9	50
10	50
平均	50

No	y
1	60
2	45
3	50
4	30
5	60
6	50
7	40
8	45
9	70
10	50
平均	50

No	z
1	20
2	45
3	50
4	95
5	80
6	55
7	5
8	15
9	50
10	85
平均	50

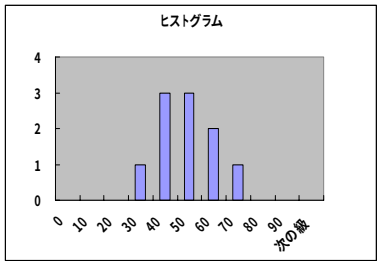
左端の資料の分布

No	x
1	50
2	50
3	50
4	50
5	50
6	50
7	50
8	50
9	50
10	50
平均	50



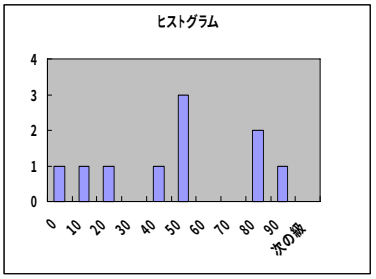
中央の資料の分布

1	60
2	45
3	50
4	30
5	60
6	50
7	40
8	45
9	70
10	50
平均	50
分散	115



右端の資料の分布

1	20
2	45
3	50
4	95
5	80
6	55
7	5
8	15
9	50
10	85
平均	50
分散	835



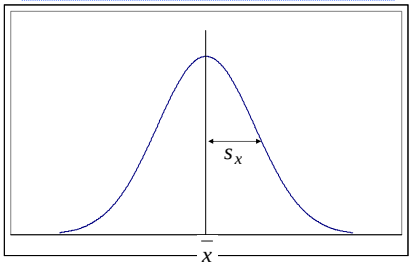
分散 s_x^2

個体名	変数 x
1	x_1
2	x_2
⋮	⋮
n	x_n

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

標準偏差 s_x

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



標準化

$$x'_i = \frac{x_i - \bar{x}}{s_x} \quad (i = 1, \cdots, n)$$

標準化された変数の平均は0，分散は1になる．（証明せよ．）

$$\overline{x'} = 0,\ s_{x'} = 1$$

ちなみに偏差値とは

i番目の個体の偏差値 $= \dfrac{x_i - \bar{x}}{s_x} \cdot 10 + 50$

No	y	偏差値
1	60	59.32505
2	45	45.33748
3	50	50
4	30	31.3499
5	60	59.32505
6	50	50
7	40	40.67495
8	45	45.33748
9	70	68.6501
10	50	50
平均	50	50

データのもつ情報量

$|x_i - \bar{x}| = i$番目のデータがもつ情報量

- もし毎日が晴れの天気であったならば，「明日は晴れる」という天気予報は何の情報もあたえない．
- 毎日，爆弾テロが起こっていては「爆弾テロが発生した」というニュースは，情報としての価値はない．
- 珍しい事ほど，あるいは，平均から離れているデータほど，情報量が大きいと考えられる．

分散 $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

はデータの平均の情報量をあらわすと考えられる．

変動

個体名	変数 x	偏差
1	x_1	$x_1 - \bar{x}$
2	x_2	$x_2 - \bar{x}$
⋮	⋮	⋮
n	x_n	$x_n - \bar{x}$

$$(x - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

はデータの総情報量をあらわすと考えられ，変動と呼ばれる．

2変数データのもつ情報量

個体名	変数 x	変数 y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
n	x_n	y_n
平均	$\bar{x}$	$\bar{y}$

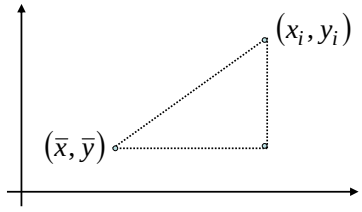
$\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}$

$= i$番目のデータがもつ情報量

2変数データのもつ情報量

$\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}$

$= i$番目のデータがもつ情報量



2変数データ全体の情報量

個体名	変数 x	変数 y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
n	x_n	y_n

データの全体の情報量 $= \sum_{i=1}^{n}\left\{ (x_i - \bar{x})^2 + (y_i - \bar{y})^2 \right\}$

$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2$

共分散 s_{xy}

個体名	変数 x	変数 y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
n	x_n	y_n

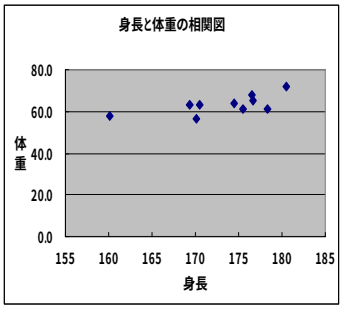
$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

社員に関する4つの調査項目

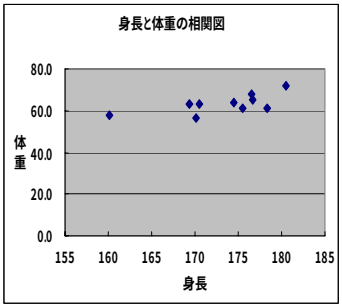
社員No	身長(x)	体重(y)	営業成績(u)	遅刻回数(v)
1	170.5	63.0	55	5
2	176.7	65.2	35	12
3	175.5	61.5	72	0
4	160.1	58.0	64	2
5	174.5	63.8	75	1
6	180.5	72.0	79	0
7	176.6	68.0	60	3
8	170.1	56.5	47	5
9	178.3	61.5	60	2
10	169.4	63.0	86	1

身長と体重の相関図（散布図）

社員No	身長(x)	体重(y)
1	170.5	63.0
2	176.7	65.2
3	175.5	61.5
4	160.1	58.0
5	174.5	63.8
6	180.5	72.0
7	176.6	68.0
8	170.1	56.5
9	178.3	61.5
10	169.4	63.0



身長と体重の相関

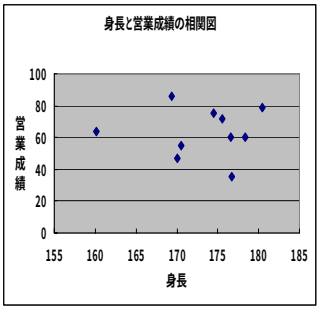


身長（x）**と体重**（y）**との間には，正の相関がある．**

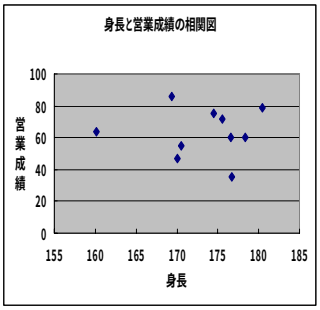
$s_{xy} = 16.6$

身長と営業成績の相関図（散布図）

社員No	身長(x)	営業成績(u)
1	170.5	55
2	176.7	35
3	175.5	72
4	160.1	64
5	174.5	75
6	180.5	79
7	176.6	60
8	170.1	47
9	178.3	60
10	169.4	86



身長と営業成績の相関

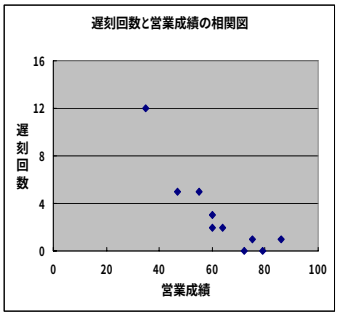


身長（x）**と営業成績**（u）**との間には，相関がない（無相関）．**

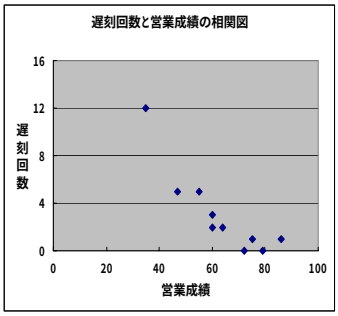
$s_{xu} = 0.02$

営業成績と遅刻回数の相関図（散布図）

社員No	営業成績(u)	遅刻回数(v)
1	55	5
2	35	12
3	72	0
4	64	2
5	75	1
6	79	0
7	60	3
8	47	5
9	60	2
10	86	1



営業成績と遅刻回数の相関



遅刻回数 (v) と営業成績 (u) との間には，**負の相関がある**．

$s_{uv} = -44.3$

相関係数 r_{xy}

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

共分散は，単位のとりかたの影響を受けるので，その大きさを単純に比較できない．

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

相関係数の性質

$$-1 \leq r_{xy} \leq 1$$

相関係数の例

	身長と体重	身長と営業成績	営業成績と遅刻回数
共分散	16.594	0.024	44.33
相関係数	0.6941	0.000	-0.888

相関係数の解釈

$$-1 \leq r_{xy} \leq 1$$

|相関係数|	意味
0〜0.2	相関はない
0.2〜0.4	ほとんど相関はない
0.4〜0.7	弱い相関がある
0.7〜1	強い相関がある

分散共分散行列

例えば，3変数 x, y, z についての分散と共分散を

$$\begin{bmatrix} s_x^2 & s_{xy} & s_{xz} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{xz} & s_{yz} & s_z^2 \end{bmatrix}$$

のように行列にまとめたものを**分散共分散行列**と呼ぶ．

相関行列

どうように，共分散の代わりに相関係数を並べたものを**相関行列**と呼ぶ．

$$\begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{bmatrix}$$

分散共分散行列も相関行列も**対称行列**である．

本日のまとめ

- 平均値，分散，標準偏差の定義，及び，それらの意味．
- 相関図，共分散，相関係数の定義，及び，それらの意味．
- 平均値，分散，標準偏差，相関図，共分散，相関係数をExcelを用いた計算．