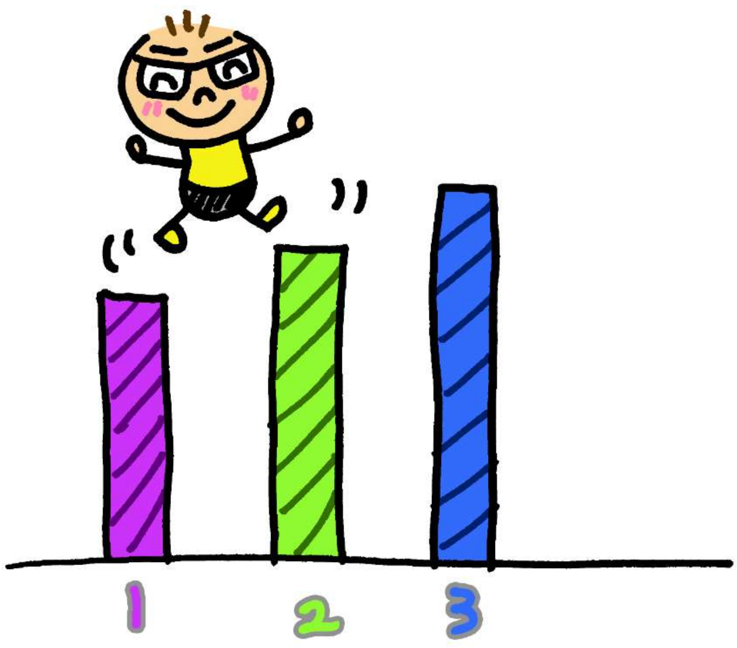
数学Ⅰ

04 データの分析



＜講義ノート＞

<u>データの分析</u>

１　データの代表値

（１）データの整理

（２）平均値

（３）中央値・最頻値

２　データの散らばり

（１）四分位数

（２）箱ひげ図

３　分散と標準偏差

（１）基本

（２）平均値と分散

（３）変数（変量）変換

４　データの相関

（１）散布図・相関表

（２）相関係数

1. データの代表値

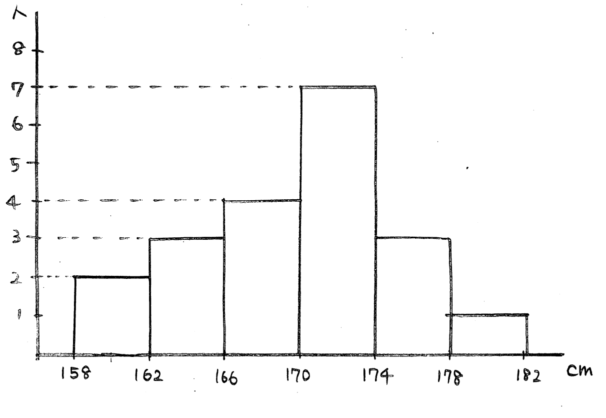
(1) データの整理

例 次の資料は、あるクラスの男子生徒の身長(cm)を調べた結果である。

176.3	161.4	178.9	172.2	167.2
172.5	170.1	167.5	168.0	175.4
169.2	165.8	174.4	163.9	158.7
165.6	170.3	173.0	170.7	171.5

階級(cm)	度数	累積度数	相対度数
158以上162未満	2	2	0.10
162 ～ 166	3	5	0.15
166 ～ 170	4	9	0.20
170 ～ 174	7	16	0.35
174 ～ 178	3	19	0.15
178 ～ 182	1	20	0.05
計	20		1.00

ヒストグラム



(2) 平均値

変量xのデータの値が$x_1, x_2, x_3, \cdots\cdots, x_n$で与えられるとき、
平均値を$\bar{x}$とすると、

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \leftarrow \frac{(\text{データの値の総和})}{(\text{データの個数})}}$$

例1 下のデータは、あるテストの生徒20人の得点である。
右の表は、これを度数分布表にまとめたものである。

59	55	52	85	66	60	65
58	91	78	45	87	18	36
74	28	41	62	76	74	

階級(点)	度数
0以上20未満	1
20 ～ 40	2
40 ～ 60	6
60 ～ 80	8
80 ～ 100	3
計	20

(1) 上の得点のデータから、得点の平均値を求めよ。

解) 得点の総和をSとすると、

$$S = 59+55+52+85+66+60+65+58+91+78+45+87 \\ +18+36+74+28+41+62+76+74$$
$$= 1210$$

よって、$\bar{x} = \frac{1210}{20} = 60.5$ 〃 (点)

(2) 右の度数分布表から、得点の平均値を求めよ。

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_r f_r}{n}$$

階級値	度数
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_r	f_r
計	n

解) $\bar{x} = \frac{10\times1 + 30\times2 + 50\times6 + 70\times8 + 90\times3}{20}$

$= 60$ 〃 (点)

<仮平均の利用>

例2 (仮平均の基本)

(1) $x_1, x_2, x_3, \cdots, x_n$ の平均値を $\bar{x}$, 仮平均を x_0 とするとき、次の等式が成り立つことを示す。

$$\bar{x} = x_0 + \frac{1}{n}\{(x_1 - x_0) + (x_2 - x_0) + \cdots + (x_n - x_0)\}$$

(2) 次のデータの仮平均を40として平均値 $\bar{x}$ を計算せよ。

33, 40, 56, 53, 41, 40, 43, 49, 31, 34

解) (1) (右辺) $= x_0 + \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n - nx_0)$

$$= x_0 + \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} - \frac{1}{n} \times nx_0$$

$$= \bar{x} = (\text{左辺})$$

よって, $\bar{x} = x_0 + \frac{1}{n}\{(x_1 - x_0) + (x_2 - x_0) + \cdots + (x_n - x_0)\}$ (証明終)

(2) $\bar{x} = 40 + \frac{1}{10}(-7 + 0 + 16 + 13 + 1 + 0 + 3 + 9 - 9 - 6)$
$= 42$

例3 (仮平均の応用)

右の表は、ある都市で、4月の毎日の最高気温を調べた結果をまとめた度数分布表である。仮平均を21℃として、度数分布表から平均値を求めよ。

階級(℃)	度数
12以上14未満	2
14～16	2
16～18	4
18～20	3
20～22	10
22～24	8
24～26	1
計	30

解) $\bar{x} = x_0 + \frac{1}{30}\{(-8)\times 2 + (-6)\times 2 + (-4)\times 4 + (-2)\times 3 + 0\times 10 + 2\times 8 + 4\times 1\}$

$= 21 + \frac{1}{30} \times (-30)$

$= 20$ (℃)。

(3) 中央値・最頻値

(i) 中央値（メジアン）
すべてのデータの値を小さい順に並べたとき、中央の順位に来る値

例1 次のデータは、9人がけん垂を何回できたかを記録したものである。

3，4，4，4，7，7，8，8，9　（単位は回）

(1) このデータの中央値を求めよ。
(2) この9人に加え「5」という1人のデータを追加したとき，全10人のデータの中央値を求めよ。

解）(1) 7（回）//

(2) 3，4，4，4，5，7，7，8，8，9

$\frac{5+7}{2} = 6$（回）//

データの個数が
- 奇数$(2n+1)$のとき、小さい方から$(n+1)$番目の値
- 偶数$(2n)$のとき、小さい方からn番目と$(n+1)$番目の平均

例2 次のデータに1つの個体（変量は整数）を加えるとき、中央値はそれぞれ何通りの場合が考えられるか。

(1) 4，6，7，9，14，15，17，19

(2) 4，6，7，9，14，15，17

解）加える個体の変量をk，中央値をMとする
(1) $k \leqq 9$，$10 \leqq k \leqq 13$，$14 \leqq k$で場合分け。
(i) $k \leqq 9$のとき

$M = 9$

(ii) $10 \leqq k \leqq 13$のとき

$M = k$

(iii) $14 \leqq k$のとき

$M = 14$

(i)～(iii)より、$1 + 4 + 1 = 6$（通り）//

(2) $k \leqq 7$, $8 \leqq k \leqq 13$, $14 \leqq k$ で場合分け

(i) $k \leqq 7$ のとき

$$M = \frac{7+9}{2} = 8$$

(ii) $8 \leqq k \leqq 13$ のとき

$$M = \frac{k+9}{2}$$

(iii) $14 \leqq k$ のとき

$$M = \frac{9+14}{2} = 11.5$$

(i)～(iii)より　$1 + 6 + 1 = 8$ (通り)〃

(ii) 最頻値(モード)

データにおいて最も個数の多い値

「度数分布表」では、最も度数の大きいデータの「階級値」

例13 (度数分布表における中央値・最頻値)

右の表は、高校生10人の通学時間を調べた結果である。

通学所要時間(分)	度数(人)
10以上20未満	1
20　～　30	4
30　～　40	3
40　～　50	2

(1) 通学所要時間の中央値を求めよ。

(答) $\frac{25+35}{2} = 30$ (分)〃

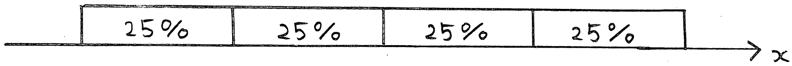
(2) 通学所要時間の最頻値を求めよ。

(答) 25 (分)〃

2. データの散らばり

(1) 四分位数

データを値の小さい方から並べて4等分する位置にくる値



5数要約を求める手順

① 最大値と最小値を求める。
② 中央値（第2四分位数）を求める。
③ 第1四分位数・第3四分位数を求める。

例 次のそれぞれの資料について，第1四分位数 Q_1，第2四分位数 Q_2，第3四分位数 Q_3 と四分位範囲，四分位偏差を求めよ。

(1) 3，5，6，7，8，8，8，9，10，11，11，11，12，13，13
(2) 3，5，6，7，8，8，8，9，10，11，11，11，12，13，13，14
(3) 3，5，6，7，8，8，8，9，10，11，11，11，12，13，13，14，15
(4) 3，5，6，7，8，8，8，9，10，11，11，11，12，13，13，14，15，15

解）

(1) 3，5，6，7，8，8，8，9，10，11，11，11，12，13，13

$Q_1 = 7$，$Q_2 = 9$，$Q_3 = 11$，$Q_3 - Q_1 = 4$，$\frac{Q_3 - Q_1}{2} = 2$ //

(2) 3，5，6，7，8，8，8，9，10，11，11，11，12，13，13，14

$Q_1 = 7.5$，$Q_2 = 9.5$，$Q_3 = 11.5$　$Q_3 - Q_1 = 4$，$\frac{Q_3 - Q_1}{2} = 2$ //

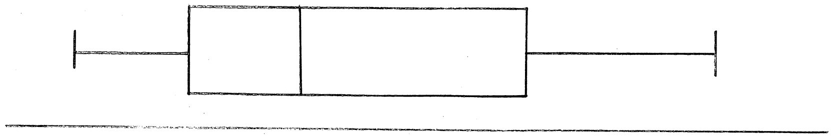
(3) 3，5，6，7，8，8，8，9，10，11，11，11，12，13，13，14，15

$Q_1 = 7.5$，$Q_2 = 10$，$Q_3 = 12.5$，$Q_3 - Q_1 = 5$，$\frac{Q_3 - Q_1}{2} = 2.5$ //

(4) 3，5，6，7，8，8，8，9，10，11，11，11，12，13，13，14，15，15

$Q_1 = 8$，$Q_2 = 10.5$，$Q_3 = 13$，$Q_3 - Q_1 = 5$，$\frac{Q_3 - Q_1}{2} = 2.5$ //

(2) 箱ひげ図



(例1) 10人の生徒の小テストの得点のデータから、5数要約をすべて求め、箱ひげ図を作成せよ。

5、7、6、5、9、7、3、4、5、8　(点)

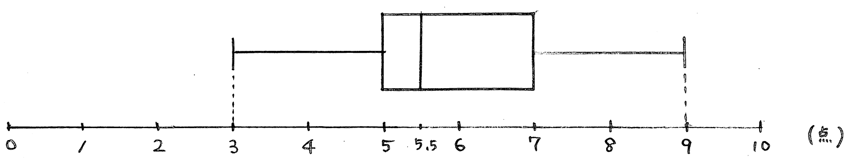
解) データを小さい順に並べると、

3、4、5、5、5、6、7、7、8、9　(点)

最小値　3(点)　　　第1四分位数　5(点)

最大値　9(点)　　　第3四分位数　7(点)

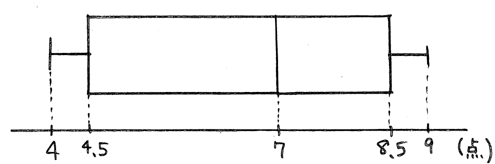
中央値 $\frac{5+6}{2} = 5.5$ (点)
(第2四分位数)



(例2) 次の箱ひげ図は9人の小テストのデータから作成したものである。次の問いに答えよ。

(1) テストの得点が1位、5位、9位の人の点数を求めよ。

(2) 四分位範囲、四分位偏差を求めよ。



解) (1) 順に、9点、7点、4点 //

(2) 第1四分位数を Q_1、第3四分位数を Q_3 とすると、

$Q_1 = 4.5$　，　$Q_3 = 8.5$

よって四分位範囲は、$Q_3 - Q_1 = 8.5 - 4.5$

$= 4$ (点) //

四分位偏差は、$\frac{Q_3 - Q_1}{2} = 2$ (点) //

3、分散と標準偏差

平均値のまわりのデータの散らばり
具合を数値で表すもの

「四分位数」
中央値のまわりのデータの散らばり
具合を数値で表すもの

(1) 基本

例1 岩崎先生が担当する、A組とB組の両方のクラスで、同じ数学のテストを実施した。その結果、どちらのクラスも平均点が70点であった。

A組	60	70	80	60	80	70	80	70	60	70	(点)
B組	100	20	50	80	90	30	60	90	100	80	(点)

確認 A組の平均：$\dfrac{60+70+80+60+80+70+80+70+60+70}{10}=\dfrac{700}{10}=70$

B組の平均：$\dfrac{100+20+50+80+90+30+60+90+100+80}{10}=\dfrac{700}{10}=70$

☆（各データ）−（平均値）＝（偏差）を調べる

$\boxed{\text{各偏差の平均}}$ A組 $\dfrac{-10+0+10-10+10+0+10+0-10+0}{10}=0$（点）

B組 $\dfrac{30-50-20+10+20-40-10+20+30+10}{10}=0$（点）

☆（偏差平方）を調べてみよう！
☆「偏差平方」の平均をとる。

分散
A組：$\dfrac{(-10)^2+0^2+10^2+(-10)^2+10^2+0^2+10^2+0^2+(-10)^2+0^2}{10}=60$（点^2）

B組：$\dfrac{30^2+(-50)^2+(-20)^2+10^2+20^2+(-40)^2+(-10)^2+20^2+30^2+10^2}{10}=740$（点^2）

$(\text{標準偏差})=\sqrt{(\text{分散})}$

① 変量xがとるn個の値$x_1, x_2, x_3, \cdots, x_n$の平均値を$\overline{x}$とする。

$x_1-\overline{x},\ x_2-\overline{x},\ x_3-\overline{x},\ \cdots,\ x_n-\overline{x}$

各データからの偏差

② 偏差の平方の平均値を「分散」といいS^2で表す。

$$S^2=\frac{1}{n}\left\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+(x_3-\overline{x})^2+\cdots+(x_n-\overline{x})^2\right\}$$

③ 分散の正の平方根を「標準偏差」といい、Sで表す。

$$S=\sqrt{\frac{1}{n}\left\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+(x_3-\overline{x})^2+\cdots+(x_n-\overline{x})^2\right\}}$$

例2 下の表は、50点満点のゲームに参加した10人の得点をまとめたものである。

番号	1	2	3	4	5	6	7	8	9	10	平均値
得点	37	44	34	35	30	41	38	33	41	37	37

(1) ゲームに参加した10人のゲームの得点について、平均値37点からの偏差の最大値は$\boxed{}$点である。

(2) 分散の値は、$\boxed{}$，標準偏差は$\boxed{}$点である。

解)

番号	1	2	3	4	5	6	7	8	9	10
得点	37	44	34	35	30	41	38	33	41	37
偏差										
(偏差)^2										

(1) 7 (点)

(2) $\frac{1}{10}(0+49+9+4+49+16+1+16+16+0)=16$

また、$\sqrt{16}=4$ (点)〃

(2) 平均値と分散

例1 標準偏差をS, データx_k ($k=1, 2, 3 \cdots n$)の平均値を$\bar{x}$, x_k^2の平均値を$\overline{x^2}$とするときの

(1) $S^2 = \overline{x^2} - (\bar{x})^2$ を示せ。

(2) (1)を用いて次のデータの標準偏差Sを求めよ。

5, 6, 8, 5, 7, 6, 5, 6, 5, 7

解)

(1)
$$S^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n}$$
$$= \frac{x_1^2 - 2x_1\bar{x} + (\bar{x})^2 + x_2^2 - 2x_2\bar{x} + (\bar{x})^2 + \cdots + x_n^2 - 2x_n\bar{x} + (\bar{x})^2}{n}$$
$$= \frac{x_1^2 + x_2^2 + \cdots + x_n^2 - 2(x_1 + x_2 + \cdots + x_n)\bar{x} + n \times (\bar{x})^2}{n}$$
$$= \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - 2 \times \frac{x_1 + x_2 + \cdots x_n}{n}\bar{x} + \frac{n(\bar{x})^2}{n}$$
$$= \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - 2(\bar{x})^2 + (\bar{x})^2$$
$$= \overline{x^2} - (\bar{x})^2 \qquad \text{(証明終)}$$

(2)
$$\bar{x} = \frac{1}{10}(5+6+8+5+7+6+5+6+5+7)$$
$$= 6$$

$$\overline{x^2} = \frac{1}{10}(5^2+6^2+8^2+5^2+7^2+6^2+5^2+6^2+5^2+7^2)$$
$$= \frac{1}{10}(25+36+64+25+49+36+25+36+25+49)$$
$$= 37$$

よって、
$$S^2 = \overline{x^2} - (\bar{x})^2$$
$$= 37 - 6^2$$
$$= 1 \qquad S = \sqrt{1} = 1 \;/\!/$$

$$\left(S^2 = \frac{1}{10}\{(-1)^2 + 0^2 + 2^2 + (-1)^2 + 1^2 + 0^2 + (-1)^2 + 0^2 + (-1)^2 + 1^2\} = 1 \;/\!/ \right)$$

例2 10個の値からなるデータがある。
そのうちの6個の平均値は3，分散は9であり，
残りの4個の平均値は8，分散は14であるとき，
この10個のデータの平均値$\overline{x}$と分散S^2を求めよ。

解) 6個の値を，$x_1, x_2, x_3, x_4, x_5, x_6$，
残りの4個の値を，x_7, x_8, x_9, x_{10} とする

$\dfrac{x_1+x_2+x_3+x_4+x_5+x_6}{6}=3$　　$\therefore x_1+x_2+x_3+x_4+x_5+x_6=18$ …①

$\dfrac{x_7+x_8+x_9+x_{10}}{4}=8$　　$\therefore x_7+x_8+x_9+x_{10}=32$ …②

①，②より

$$\overline{x}=\frac{x_1+x_2+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}}{10}$$

$$=\frac{18+32}{10}\quad(①，②より)$$

$$=5$$

また，$\frac{1}{6}(x_1^2+x_2^2+x_3^2+x_4^2+x_5^2+x_6^2)-3^2=9$ より

$x_1^2+x_2^2+x_3^2+x_4^2+x_5^2+x_6^2=18\times 6=108$ …③

さらに $\frac{1}{4}(x_7^2+x_8^2+x_9^2+x_{10}^2)-8^2=14$ より

$x_7^2+x_8^2+x_9^2+x_{10}^2=78\times 4=312$ …④

③④から

$$\overline{x^2}=\frac{x_1^2+x_2^2+x_3^2+x_4^2+x_5^2+x_6^2+x_7^2+x_8^2+x_9^2+x_{10}^2}{10}$$

$$=\frac{420}{10}=42$$

したがって，

$$S^2=\overline{x^2}-(\overline{x})^2$$

$$=42-5^2=17$$

(3) 変数変換

準備（仮平均と変数変換）

例1 A，B，C，D，Eの5人があるゲームをし，その得点を以下の表にまとめた。

参加者	A	B	C	D	E
得点（点）	6500	8500	7000	5500	9500

この5人の得点の平均値を求めよ。

解) $\boxed{\text{ふつう}}$ $\dfrac{6500+8500+7000+5500+9500}{5}=7400$（点）

$\boxed{\text{仮平均を利用}}$ 仮平均を7000点とすれば，

A	B	C	D	E
−500	1500	0	−1500	2500

これらの平均値 $\dfrac{-500+1500+0-1500+2500}{5}=400$

よって，$7000+400=7400$（点）//

$\boxed{\text{さらに工夫}}$

A	B	C	D	E
−5	15	0	−15	25

これらの平均値 $\dfrac{-5+15+0-15+25}{5}=4$

よって，$7000+4\times100=7400$（点）//

例2 (1) $x_i\ (i=1,2,3,\cdots\cdots,n)$ の平均値を $\overline{x}$，また，$x_i=au_i+b$ とおき $u_i\ (i=1,2,3,\cdots\cdots,n)$ の平均値を $\overline{u}$ とするとき，

$$\overline{x}=a\overline{u}+b$$

となることを示せ。

(2) 次のデータの平均値 $\overline{x}$ を，$x=5u+100$ として求めよ。

90，100，80，110，140，115，125，130，95，85

解)

(1) $x_1=au_1+b,\ x_2=au_2+b,\ \cdots\cdots,\ x_n=au_n+b$ から

$$\begin{aligned}\overline{x}&=\frac{1}{n}(x_1+x_2+\cdots\cdots+x_n)\\&=\frac{1}{n}\{(au_1+b)+(au_2+b)+\cdots\cdots+(au_n+b)\}\\&=\frac{1}{n}\{a(u_1+u_2+\cdots\cdots+u_n)+nb\}\\&=a\times\frac{1}{n}(u_1+u_2+\cdots\cdots+u_n)+b\\&=a\overline{u}+b\end{aligned}$$

（証明終）

(2) uの平均値を$\overline{u}$とすると，表から

x	90	100	80	110	140	115	125	130	95	85
$x-100$	-10	0	-20	10	40	15	25	30	-5	-15
u	-2	0	-4	2	8	3	5	6	-1	-3

$$\overline{u}=\frac{1}{10}(-2+0-4+2+8+3+5+6-1-3)=1.4$$

よって $\overline{x}=5\times 1.4+100=107$ //

例3 (1) $x=au+b$で，uの標準偏差がs'のとき，xの標準偏差sは，
$s=|a|s'$で表されることを示せ。

(2) 次のデータの標準偏差sを$x=5u+100$として求めよ。

90, 100, 80, 110, 140, 115, 125, 130, 95, 85

解) (1) $x_1, x_2, \cdots, x_n$の平均値を$\overline{x}$とすると，$\overline{x}=a\overline{u}+b$ だから，

$$s^2=\frac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}$$

$$=\frac{1}{n}[\{(au_1+b)-(a\overline{u}+b)\}^2+\{(au_2+b)-(a\overline{u}+b)\}^2+\cdots$$
$$\cdots+\{(au_n+b)-(a\overline{u}+b)\}^2]$$

$$=\frac{1}{n}[\{a(u_1-\overline{u})\}^2+\{a(u_2-\overline{u})\}^2+\cdots+\{a(u_n-\overline{u})\}^2]$$

$$=\frac{1}{n}\{(u_1-\overline{u})^2+(u_2-\overline{u})^2+\cdots+(u_n-\overline{u})^2\}\times a^2$$

$$=a^2s'^2$$

よって $s=|a|s'$　　　(証明終)

(2) $x=5u+100$ だから

x	90	100	80	110	140	115	125	130	95	85
$x-100$	-10	0	-20	10	40	15	25	30	-5	-15
u	-2	0	-4	2	8	3	5	6	-1	-3

uの標準偏差s'の2乗は，

$$s'^2=\frac{1}{10}\{(-2)^2+0^2+(-4)^2+2^2+8^2+3^2+5^2+6^2+(-1)^2+(-3)^2\}$$
$$-\{\frac{1}{10}(-2+0-4+2+8+3+5+6-1-3)\}^2$$

$$=\frac{1}{10}\times 168-\left(\frac{14}{10}\right)^2$$

$$=\frac{1680-196}{10}=14.84$$

よって $s=5\times\sqrt{14.84}=\sqrt{371}\fallingdotseq 19.3$ //

〈度数分布表と分散・標準偏差〉

変量xのデータが右のような度数分布表で与えられているとき，分散S^2は，

$$S^2=\frac{(x_1-\bar{x})^2f_1+(x_2-\bar{x})^2f_2+\cdots+(x_r-\bar{x})f_r}{n}$$
$$=\frac{x_1^2f_1+x_2^2f_2+\cdots+x_r^2f_r}{n}-(\bar{x})^2$$

→ 標準偏差 $S=\sqrt{S^2}$

階級値x	度数f
x_1	f_1
x_2	f_2
⋮	⋮
x_r	f_r
計	n

例 右の度数分布表で与えられたデータの平均値$\bar{x}$と分散・標準偏差を求めよ。

階級	度数
120～130	2
130～140	3
140～150	10
150～160	3
160～170	2
計	20

解)

階級値x	度数f	xf	$x-\bar{x}$	$(x-\bar{x})^2f$	x^2f
125	2	250	−20	800	31250
135	3	405	−10	300	54675
145	10	1450	0	0	210250
155	3	465	10	300	72075
165	2	330	20	800	54450
計	20	2900	0	2200	422700

$\bar{x}=\frac{2900}{20}=145$〃　$S^2=\frac{2200}{20}=110$〃　$S=\sqrt{110}\fallingdotseq 10.5$〃

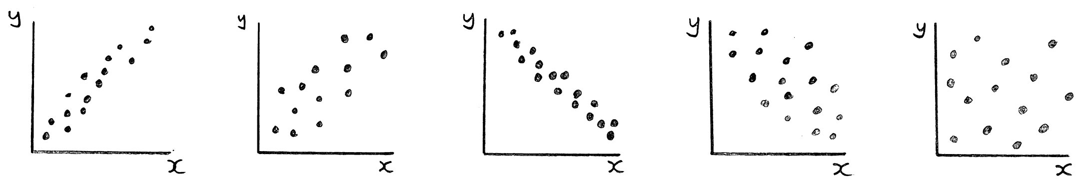
(別解) $S^2=21135-145^2=110$〃

4. データの相関

(1) 散布図・相関表

(i) 散布図

2つの変量からなるデータにおいて、これらの値を座標とする点を平面上にとり、2変量の関係を示した図。



(ii) 相関表

2つの変量の相関関係について、2つの度数分布表を組み合わせたもの

例2 次の表は、生徒15人に対して行った数学と化学の得点を示したものである。数学の得点を x (点)、化学の得点を y (点) として、x と y の相関表を作れ。

数学	81	69	73	92	84	52	75	60	67	70	63	58	76	48	85
化学	89	72	75	87	81	59	81	53	62	69	57	51	65	54	94

解)

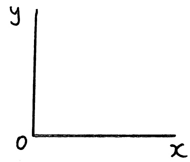
y \ x	40~50	50~60	60~70	70~80	80~90	90~100	計
90~100							
80~90							
70~80							
60~70							
50~60							
計							

注「相関関係」と「因果関係」は違う!

相関関係	因果関係
2つの変量の間で一方が増加するにつれて他方が増加または、減少する関係。	いくつかの事柄の関係において、一方が原因で他方が結果であるというつながりがあること。

(2) 相関係数

準備（共分散）



	$x_k-\overline{x}$	$y_k-\overline{y}$	$(x_k-\overline{x})(y_k-\overline{y})$
領域A			
領域B			
領域C			
領域D			

2つの変量 x、y について n 個のデータ

(x_1, y_1), (x_2, y_2), …, (x_n, y_n) を考える。

x、y の平均値をそれぞれ $\overline{x}$、$\overline{y}$ としその偏差積

$(x_1-\overline{x})(y_1-\overline{y})$, $(x_2-\overline{x})(y_2-\overline{y})$, …, $(x_n-\overline{x})(y_n-\overline{y})$ の符号

を調べる。

共分散 S_{xy}

$$S_{xy}=\frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{n}$$

例1 下の表は、5人の一日の勉強時間(h)と小テストの点数(点)を調べた結果である。

x	3	7	5	1	4
y	6	10	8	4	7

共分散 S_{xy} を求めることにより、相関について調べよ。

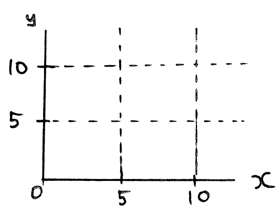
解）

$$\overline{x}=\frac{1}{5}(3+7+5+1+4)=4$$

$$\overline{y}=\frac{1}{5}(6+10+8+4+7)=7$$

$$S_{xy}=\frac{(-1)\times(-1)+3\times3+1\times1+(-3)\times(-3)+0\times0}{5}$$

$$=4\ (>0)$$



$$S_x=\sqrt{\frac{1}{5}\{(-1)^2+3^2+1^2+(-3)^2+0^2\}}=2$$

$$S_y=\sqrt{\frac{1}{5}\{(-1)^2+3^2+1^2+(-3)^2+0^2\}}=2$$

$$r=\frac{S_{xy}}{S_xS_y}=\frac{4}{2\times2}=1$$

例2 下の表は、5匹の犬の体重 x(kg) と体長 y(cm) を測った結果である。

x	6	25	15	4	40
y	35	60	45	25	70

共分散 S_{xy} を求めることにより、相関について調べよ。

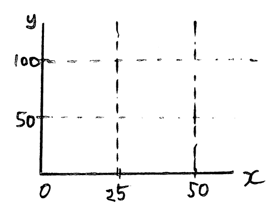
解）

$$\overline{x}=\frac{1}{5}(6+25+15+4+40)=18$$

$$\overline{y}=\frac{1}{5}(35+60+45+25+70)=47$$

$$S_{xy}=\frac{(-12)\times(-12)+7\times13+(-3)\times(-2)+(-14)\times(-22)+22\times23}{5}$$

$$=211\ (>0)$$



$$S_x=\sqrt{\frac{1}{5}\{(-12)^2+7^2+(-3)^2+(-14)^2+22^2\}}=\frac{21\sqrt{2}}{\sqrt{5}}$$

$$S_y=\sqrt{\frac{1}{5}\{(-12)^2+13^2+(-2)^2+(-22)^2+23^2\}}=\sqrt{266}$$

$$r=\frac{S_{xy}}{S_xS_y}=\frac{211}{\frac{21\sqrt{2}}{\sqrt{5}}\times\sqrt{266}}=\frac{211\sqrt{5}}{42\sqrt{133}}$$

$$\fallingdotseq 0.97$$

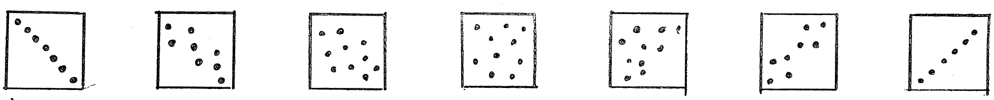
$$r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\frac{1}{n}\{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})\}}{\sqrt{\frac{1}{n}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2\}}\sqrt{\frac{1}{n}\{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2\}}}$$

共分散S_{xy}を標準偏差S_xとS_yの積$S_x \cdot S_y$で割った値

$$= \frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}\sqrt{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2}}$$

$-1 \leqq r \leqq 1$ となる

rが1に近い：
rが-1に近い：



$r=-1$ $r=-0.8$ $r=-0.5$ $r=0$ $r=0.5$ $r=0.8$ $r=1$

<$-1 \leqq r \leqq 1$ になることの証明>

$$相関係数 r = \frac{S_{xy}}{S_x S_y}$$

$$= \frac{\frac{1}{n}\{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots\cdots+(x_n-\bar{x})(y_n-\bar{y})\}}{\sqrt{\frac{1}{n}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}\sqrt{\frac{1}{n}\{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2\}}}$$

$$= \frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}\sqrt{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2}}$$

解答

(1) $F=(a_1^2+a_2^2+\cdots+a_n^2)t^2+2(a_1b_1+a_2b_2+\cdots+a_nb_n)t+(b_1^2+b_2^2+\cdots+b_n^2)$

$=a_1^2t^2+2a_1b_1t+b_1^2+a_2^2t^2+2a_2b_2t+b_2^2+\cdots+a_n^2t^2+2a_nb_nt+b_n^2$

$=(a_1t+b_1)^2+(a_2t+b_2)^2+\cdots+(a_nt+b_n)^2 \geqq 0$ (証明終)

(2)(i) $a_1^2+a_2^2+\cdots+a_n^2=0$, つまり, $a_1=a_2=\cdots=a_n=0$ のとき

$(a_1b_1+a_2b_2+\cdots+a_nb_n)^2=(a_1^2+a_2^2+\cdots+a_n^2)(b_1^2+b_2^2+\cdots b_n^2)$

は成立。

(ii) $a_1^2+a_2^2+\cdots+a_n^2>0$ のとき

2次方程式

$(a_1^2+a_2^2+\cdots+a_n^2)t^2+2(a_1b_1+a_2b_2+\cdots+a_nb_n)t+b_1^2+b_2^2+\cdots+b_n^2=0$

の判別式を D とすると, $D \leqq 0$

$D/4=(a_1b_1+a_2b_2+\cdots+a_nb_n)^2-(a_1^2+a_2^2+\cdots+a_n^2)(b_1^2+b_2^2+\cdots+b_n^2) \leqq 0$

$\therefore (a_1b_1+a_2b_2+\cdots+a_nb_n)^2 \leqq (a_1^2+a_2^2+\cdots+a_n^2)(b_1^2+b_2^2+\cdots+b_n^2)$

(i)(ii)より, $(a_1b_1+a_2b_2+\cdots+a_nb_n)^2 \leqq (a_1^2+a_2^2+\cdots+a_n)(b_1^2+b_2^2+\cdots+b_n^2)$

(証明終)

(3) (2)より $(a_1b_1+a_2b_2+\cdots+a_nb_n)^2 \leqq (a_1^2+a_2^2+\cdots+a_n^2)(b_1^2+b_2^2+\cdots+b_n^2)$

題意より, $a_1^2+a_2^2+\cdots+a_n^2>0$, $b_1^2+b_2^2+\cdots+b_n^2>0$ だから

$$\frac{(a_1b_1+a_2b_2+\cdots+a_nb_n)^2}{(a_1^2+a_2^2+\cdots+a_n^2)(b_1^2+b_2^2+\cdots+b_n^2)} \leqq 1$$

$$\Leftrightarrow -1 \leqq \frac{a_1b_1+a_2b_2+\cdots+a_nb_n}{\sqrt{a_1^2+a_2^2+\cdots+a_n^2}\sqrt{b_1^2+b_2^2+\cdots+b_n^2}} \leqq 1$$

(証明終)

(4) (3)の結果から

$$-1 \leqq \frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}\sqrt{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2}}$$

$\Leftrightarrow -1 \leqq r \leqq 1$ (証明終)