

データの分析

1 データの整理

右の度数分布表は、A 高校の 20 人について、1 日にみたテレビの時間を記入したものである。次の問いに答えよ。

- (1) テレビをみた時間が 85 分未満の生徒は何人いるか。
- (2) テレビをみた時間が 95 分以上の生徒は全体の何%であるか。
- (3) 右の度数分布表をもとにして、ヒストグラムをかけ。

階級 (分)	階級値 (分)	度数 (人)	相対度数
55 以上～65 未満	60	2	0.10
65 ～75	70	2	0.10
75 ～85	80	3	0.15
85 ～95	90	4	0.20
95 ～105	100	6	0.30
105 ～115	110	2	0.10
115 ～125	120	1	0.05
合計		20	1.00

要 点

ある集団を構成する人や物の特性を表す数量を **変量** といい、変量の個々の値や、その集まりのことを **データ** という。

データを整理するとき、設定した各区間を **階級** といい、各階級の中央の値をその階級の **階級値** という。また、各階級に入る値の個数をその階級の **度数** といい、各階級に度数を対応させた表を **度数分布表** という。データ全体に対する各階級の度数の割合を、その階級の **相対度数** という。縦軸に度数、横軸に階級をとったグラフを **ヒストグラム** という。

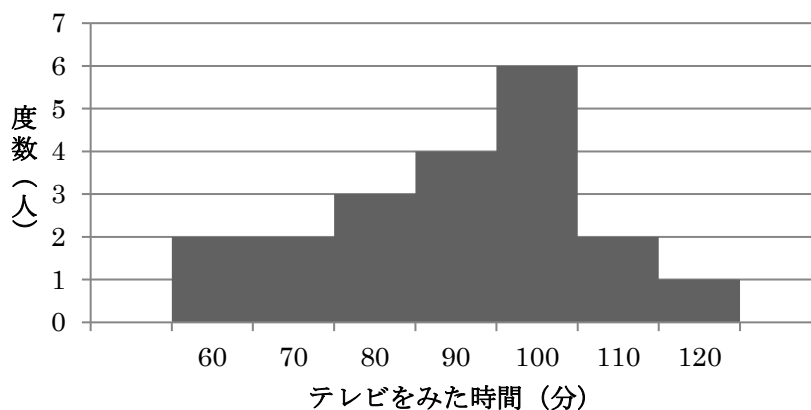
解答

(1) $2+2+3=7$ (人)

(2) $0.30+0.10+0.05=0.45$

したがって **45%**

(3)



2 平均値

(1) 次のデータは、ある高校生7人が1ヵ月にカレーライスを食べた回数 x を調べたものである。

10, 8, 4, 6, 9, 5, 7 (回)

このデータの平均値 \bar{x} を求めよ。

(2) 右の表から、テレビをみた時間 x の平均値を求めよ。

階級 (分)	階級値 x (分)	度数 f (人)
55 以上～65 未満	60	2
65 ～75	70	2
75 ～85	80	3
85 ～95	90	4
95 ～105	100	6
105 ～115	110	2
115 ～125	120	1
合計		20

要 点

平均値

変量 x の n 個の値 x_1, x_2, \dots, x_n からなるデータについて、値の合計を個数 n で割った値を **平均値** といい、記号 \bar{x} で表す。

$$\text{平均値} = \frac{\text{変量の値の合計}}{\text{変量の値の個数}}, \quad \bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$$

度数分布表からの平均値

右の度数分布表では、 x_1 が f_1 個、 x_2 が f_2 個、 \dots 、 x_r が f_r 個あるとみて平均値を計算する。

$$\text{平均値} = \frac{\text{(変量の値} \times \text{度数) の合計}}{\text{変量の値の個数}}$$

$$\bar{x} = \frac{1}{n}(x_1 f_1 + x_2 f_2 + \dots + x_r f_r)$$

ただし $n = f_1 + f_2 + \dots + f_r$

階級値 x	度数 f
x_1	f_1
x_2	f_2
\vdots	\vdots
x_r	f_r
合計	n

解答

(1) $\bar{x} = \frac{1}{7}(10+8+4+6+9+5+7) = \frac{49}{7} = 7$ (回)

(2) $\bar{x} = \frac{1}{20}(60 \times 2 + 70 \times 2 + 80 \times 3 + 90 \times 4 + 100 \times 6 + 110 \times 2 + 120 \times 1)$
 $= \frac{1}{20} \times 1800 = 90$ (分)

3 中央値, 最頻値

次のデータは, ある高校生 8 人が 1 ヶ月に読んだ本の冊数である。ただし, 教科書, 参考書, 雑誌, 漫画は除く。

3, 2, 0, 1, 3, 1, 1, 2 (冊)

- (1) このデータの中央値を求めよ。
- (2) このデータ的最頻値を求めよ。

要 点

中央値

データの値を大きさの順に並べたとき, 中央の順位にくる値を **中央値** または **メジアン** という。データの値の個数が偶数のときは, 中央に並ぶ 2 つの値の平均値を中央値とする。

最頻値

データに最も多く現れる値を **最頻値** または **モード** という。

解答

- (1) 小さい方から順に並べると 0, 1, 1, 1, 2, 2, 3, 3

これより, 中央値は $\frac{1+2}{2}=1.5$ (冊)

- (2) 最頻値は 1 (冊)

4 範囲, 四分位数, 四分位範囲, 四分位偏差

次のデータは, A 社の従業員 11 人の年収を調べたものである。

490, 470, 540, 520, 500, 480, 490, 550, 460, 470, 530 (万円)

次の問いに答えよ。

- (1) このデータの範囲を求めよ。
- (2) このデータの四分位数 Q_1 , Q_2 , Q_3 を求めよ。
- (3) このデータの四分位範囲と四分位偏差を求めよ。

要 点

範囲

データの最大値から最小値を引いた値を **範囲** という。

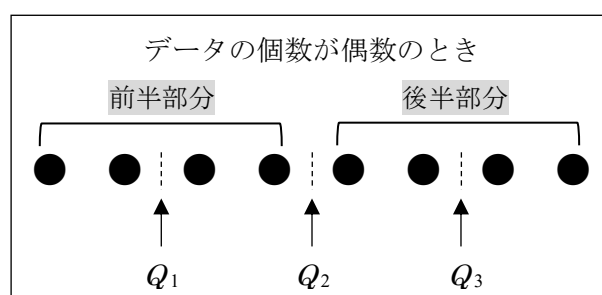
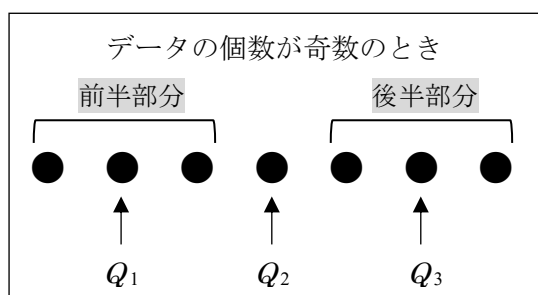
$$\text{範囲} = \text{最大値} - \text{最小値}$$

四分位数

データの値を小さい方から順に並び、中央値によって前半部分と後半部分の2つに分ける。データの値の個数が奇数のときは、中央値を1つ除いてから、前半部分と後半部分を考える。

最小値を含む前半部分の中央値を **第1四分位数**，データ全体の中央値を **第2四分位数**，最大値を含む後半部分の中央値を **第3四分位数** といい、それぞれ Q_1 ， Q_2 ， Q_3 で表す。

これらをまとめて **四分位数** という。



四分位範囲，四分位偏差

第3四分位数 Q_3 から第1四分位数 Q_1 を引いた値を **四分位範囲** という。

また、四分位範囲を2で割った値を **四分位偏差** という。

$$\text{四分位範囲} = Q_3 - Q_1, \quad \text{四分位偏差} = \frac{Q_3 - Q_1}{2}$$

解答

(1) 最大値は550万円，最小値は460万円であるから，範囲は $550 - 460 = 90$ (万円)

(2) 小さい方から順に並べると 460, 470, 470, 480, 490, 490, 500, 520, 530, 540, 550

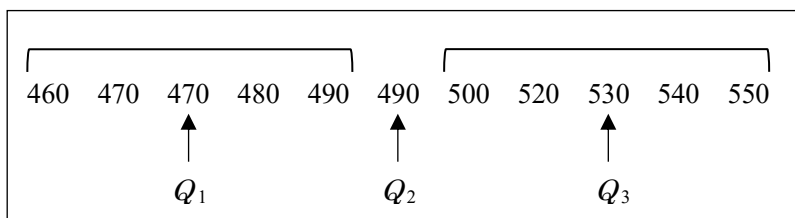
中央値から $Q_2 = 490$ (万円)

前半部分の中央値から

$$Q_1 = 470 \text{ (万円)}$$

後半部分の中央値から

$$Q_3 = 530 \text{ (万円)}$$



(3) $Q_1 = 470$ ， $Q_3 = 530$ であるから

$$\text{四分位範囲は } 530 - 470 = 60 \text{ (万円)}$$

$$\text{四分位偏差は } \frac{60}{2} = 30 \text{ (万円)}$$

5 箱ひげ図

次のデータは、A社の従業員11人、B社の従業員9人の年収を調べたものである。それぞれの箱ひげ図をかき、散らばりの度合いを比較せよ。

A社：490, 470, 540, 520, 500, 480, 490, 550, 460, 470, 530 (万円)

B社：390, 350, 370, 360, 680, 900, 400, 350, 700 (万円)

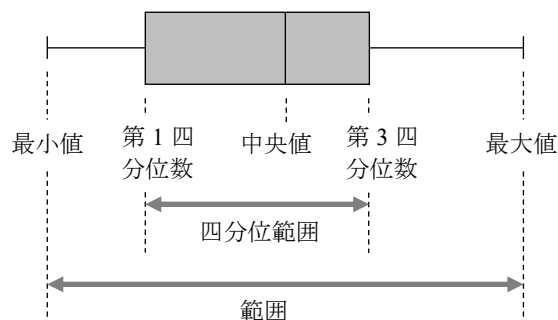
要 点

箱ひげ図

最小値, 第1四分位数, 中央値 (第2四分位数), 第3四分位数, 最大値を, 中央値で仕切られた長方形の箱と, その両端から伸びるひげのような線で表した図を **箱ひげ図** という。

箱ひげ図から, 範囲や四分位範囲を読み取ることができる。

〈注意〉 範囲や四分位範囲が小さいほど, データの値は中央値の近くに集中し, 散らばりの度合いは小さいと考えられる。



解答

A社の最小値, Q_1 , Q_2 , Q_3 , 最大値は, **4** から 460, 470, 490, 530, 550 (万円)

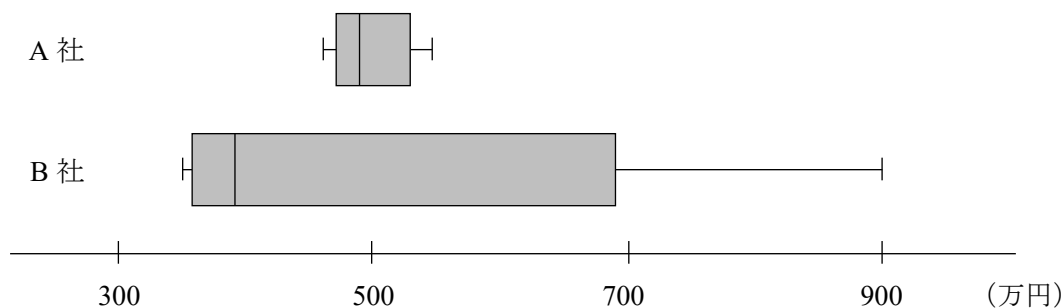
B社の最小値, Q_1 , Q_2 , Q_3 , 最大値を求める。

小さい方から順に並べると 350, 350, 360, 370, 390, 400, 680, 700, 900

これから, 最小値, 最大値は 350, 900 (万円)

また $Q_2=390$ (万円) $Q_1 = \frac{350+360}{2} = 355$ (万円) $Q_3 = \frac{680+700}{2} = 690$ (万円)

以上から, A社とB社の箱ひげ図は次のようになる。



箱ひげ図から読み取れる範囲や四分位範囲から, B社よりもA社の方が散らばりの度合いが小さい。

6 外れ値

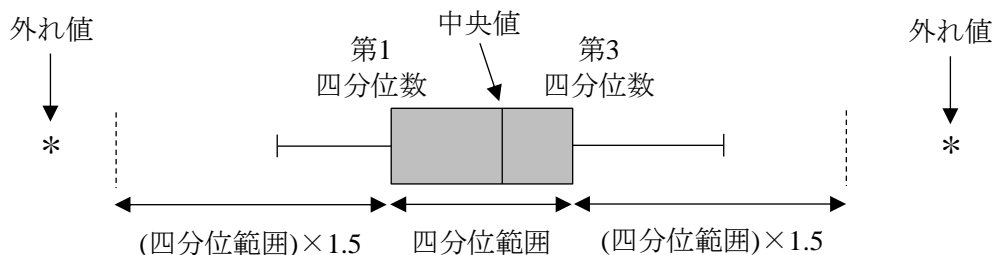
次のデータは、C社の従業員10人の年収を調べたものである。外れ値があれば求めて、箱ひげ図をかけ。

C社：350, 400, 410, 470, 500, 520, 530, 550, 600, 1500（万円）

要 点

外れ値

データの中に、他の値から極端にかけ離れた値が含まれることがある。その値を **外れ値** という。外れ値は、箱ひげ図の箱の両端から四分位範囲の1.5倍よりも外側に離れている値である。箱ひげ図では外れ値を除いた最大値および最小値までひげを引き、外れ値はひげの外に「*」などでかくことがある。



- 〈注意〉・四分位数は、外れ値を除かないすべてのデータから求め、その値にもとづいて箱ひげ図の箱をかく。
- ・測定ミス、入力ミスなど原因が分かっているものは「異常値」とよび、外れ値と区別する。
 - ・外れ値の背景を探ることで、問題解決の手がかりが見つかる場合もある。例えば、年収1500万円の従業員の心がまえを聞くことで、年収アップにつながるヒントが得られるかもしれない。
 - ・平均値は外れ値による影響を受けやすいが、中央値は外れ値の影響をあまり受けない。

解答

中央値(第2四分位数 Q_2)は $\frac{500 + 520}{2} = 510$ (万円)

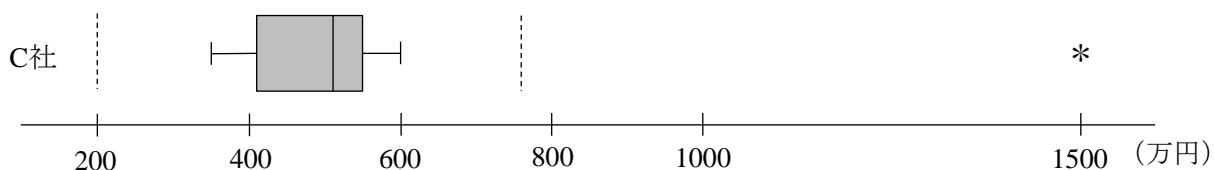
第1四分位数 Q_1 は 410 (万円) 第3四分位数 Q_3 は 550 (万円)

また、四分位範囲は $550 - 410 = 140$ (万円) であるから $(\text{四分位範囲}) \times 1.5 = 140 \times 1.5 = 210$

よって、箱の両端から210離れた値、すなわち200や760より外側にある値は外れ値である。

したがって、**外れ値は1500(万円)**

箱ひげ図は次の図のようになる。



7 分散

次のデータは、ある高校生7人が1カ月にカレーライスを食べた回数 x を調べたものである。

10, 8, 4, 6, 9, 5, 7 (回)

このデータの分散 s^2 を求めよ。

要 点**分散**

変量 x の n 個の値 x_1, x_2, \dots, x_n の平均値を \bar{x} とするとき、 $x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$ をそれぞれの値の **偏差** という。偏差の2乗の平均値を、変量 x の **分散** といい、 s^2 で表す。

$$\text{分散} = (\text{偏差})^2 \text{の平均値}, \quad s^2 = \frac{1}{n} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2\}$$

〈注意〉 x^2 の平均値を $\overline{x^2}$ で表すとき、分散 s^2 は次のようにも表される。

$$\text{分散} = (\overline{x^2 \text{の平均値}}) - (x \text{の平均値})^2, \quad s^2 = \overline{x^2} - (\bar{x})^2$$

このことは、 $\bar{x} = m$ とおいて、次のように確かめることができる。

$$\begin{aligned} s^2 &= \frac{1}{n} \{(x_1 - m)^2 + (x_2 - m)^2 + \dots + (x_n - m)^2\} \\ &= \frac{1}{n} \{(x_1^2 - 2x_1m + m^2) + (x_2^2 - 2x_2m + m^2) + \dots + (x_n^2 - 2x_nm + m^2)\} \\ &= \frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2 - 2x_1m - 2x_2m - \dots - 2x_nm + m^2 + m^2 + \dots + m^2) \\ &= \frac{1}{n} \{(x_1^2 + x_2^2 + \dots + x_n^2) - 2m(x_1 + x_2 + \dots + x_n) + n \cdot m^2\} \\ &= \frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2) - 2m \cdot \frac{1}{n} (x_1 + x_2 + \dots + x_n) + m^2 \\ &= \frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2) - 2m \cdot m + m^2 = \frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2) - m^2 \end{aligned}$$

解答

$$\text{平均値は } \bar{x} = \frac{1}{7} (10 + 8 + 4 + 6 + 9 + 5 + 7) = \frac{49}{7} = 7 \text{ (回)}$$

$$\text{偏差は } 3, 1, -3, -1, 2, -2, 0 \text{ (回)}$$

$$\text{よって、分散は } s^2 = \frac{1}{7} \{3^2 + 1^2 + (-3)^2 + (-1)^2 + 2^2 + (-2)^2 + 0^2\} = \frac{28}{7} = 4$$

$$\text{別解} \quad \text{平均値は } \bar{x} = 7 \text{ (回)} \quad \overline{x^2} = \frac{1}{7} (10^2 + 8^2 + 4^2 + 6^2 + 9^2 + 5^2 + 7^2) = \frac{371}{7} = 53$$

$$\text{したがって } s^2 = 53 - 7^2 = 4$$

8 標準偏差

次のデータは、ある高校生7人が1カ月に読んだ本の冊数 x である。ただし、教科書、参考書、雑誌、漫画は除く。

4, 2, 1, 1, 4, 2, 0 (冊)

このデータの標準偏差 s を求めよ。ただし、 $\sqrt{2} = 1.4$ とする。

要 点

標準偏差

分散の正の平方根を **標準偏差** といい、 s で表す。

$$\text{標準偏差} = \sqrt{\text{分散}}, \quad s = \sqrt{\frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2\}}$$

〈注意〉 x^2 の平均値を $\overline{x^2}$ で表すとき、標準偏差 s は次のようにも表される。

$$\text{標準偏差} = \sqrt{(\overline{x^2} - (\bar{x})^2)}, \quad s = \sqrt{\overline{x^2} - (\bar{x})^2}$$

解答

平均値は $\bar{x} = \frac{1}{7}(4+2+1+1+4+2+0) = \frac{14}{7} = 2$ (冊)

偏差は 2, 0, -1, -1, 2, 0, -2 (冊)

よって、標準偏差は $s = \sqrt{\frac{1}{7}\{2^2+0^2+(-1)^2+(-1)^2+2^2+0^2+(-2)^2\}} = \sqrt{\frac{14}{7}} = \sqrt{2} = 1.4$ (冊)

別解 平均値は $\bar{x} = 2$ (冊) $\overline{x^2} = \frac{1}{7}(4^2+2^2+1^2+1^2+4^2+2^2+0^2) = \frac{42}{7} = 6$

したがって $s = \sqrt{6-2^2} = \sqrt{2} = 1.4$ (冊)

9 散布図

右のデータは、ある高校生7人が1カ月にカレーライスを食べた回数 x と、1カ月に読んだ本の冊数 y を調べたものである。ただし、 y は教科書、参考書、雑誌、漫画を除く。

高校生	A	B	C	D	E	F	G
カレーライス (回)	10	8	4	6	9	5	7
本 (冊)	4	2	1	1	4	2	0

カレーライスを食べた回数 x を横軸、読んだ本の冊数 y を縦軸として散布図をかけ。また、 x と y の間には、どのような相関関係があるといえるか。

要 点

散布図

2つの変量の値の組を座標平面上の点で表したものを **散布図** という。

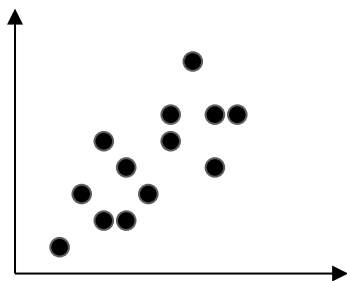
散布図と相関関係

2つの変量 x, y について、

一方の値が大きくなると他方の値も大きくなる傾向があるとき、 x と y の間には

正の相関関係がある

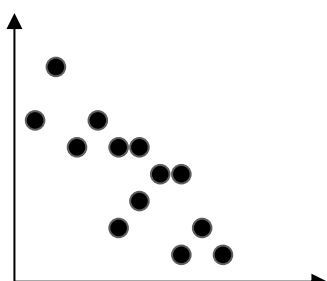
という。



一方の値が大きくなると他方の値は小さくなる傾向があるとき、 x と y の間には

負の相関関係がある

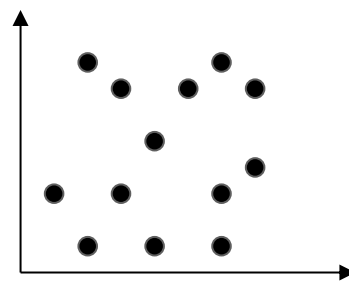
という。



正、負いずれの相関関係も見られないとき、 x と y の間には

相関関係がない

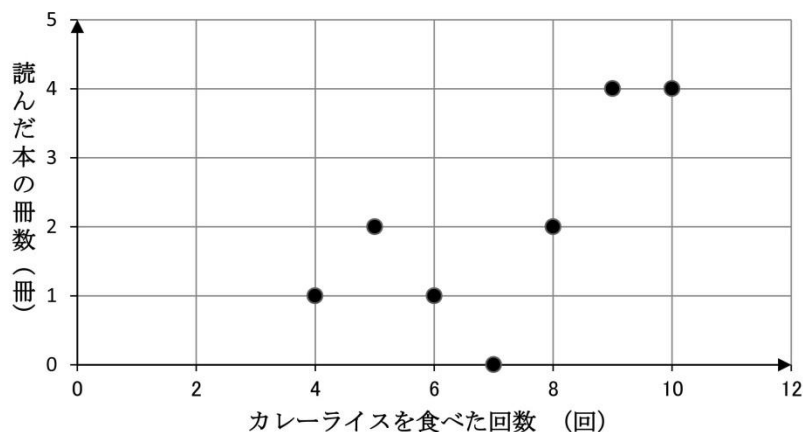
という。



解答

散布図は右のようになる。

右の散布図から、 x と y の間には **正の相関関係がある** といえる。



10 相関係数

右のデータは、ある高校生7人が1カ月にカレーライスを食べた回数 x と、1カ月に読んだ本の冊数 y を調べたものである。ただし、 y は教科書、参考書、雑誌、漫画を除く。

高校生	A	B	C	D	E	F	G
カレーライス (回)	10	8	4	6	9	5	7
本 (冊)	4	2	1	1	4	2	0

x と y の相関係数 r を求めよ。ただし、 $\sqrt{2} = 1.4$ とする。また、 x と y の間には、どのような相関関係があるといえるか。

要 点

共分散

偏差の積 $(x - \bar{x})(y - \bar{y})$ の平均値を、 x と y の **共分散** といい、 s_{xy} で表す。

共分散 = 偏差の積の平均値

$$s_{xy} = \frac{1}{n} \{ (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}) \}$$

相関係数

x の標準偏差 s_x と y の標準偏差 s_y の積 $s_x s_y$ で、共分散 s_{xy} を割った値を **相関係数** といい、 r で表す。

$$\text{相関係数} = \frac{x \text{ と } y \text{ の共分散}}{(x \text{ の標準偏差}) \times (y \text{ の標準偏差})}, \quad r = \frac{s_{xy}}{s_x s_y}$$

分母と分子に n を掛けると、次の式が得られる。

$$\text{相関係数} = \frac{(x - \bar{x})(y - \bar{y}) \text{ の合計}}{\sqrt{(x - \bar{x})^2 \text{ の合計}} \sqrt{(y - \bar{y})^2 \text{ の合計}}}$$

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

相関係数 r のとり得る値の範囲は $-1 \leq r \leq 1$ であることが知られている。 r の値から、2 つの変量には次のような相関関係があるといえる。

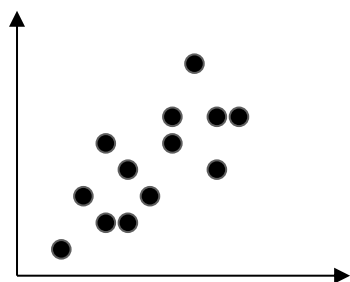
r が正のとき、正の相関関係がある。

r が 1 に近い値であるほど、正の相関関係が強い。

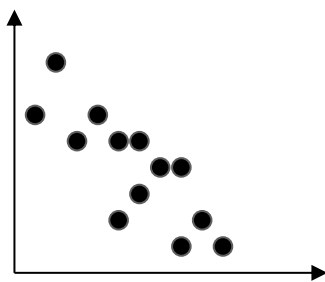
r が負のとき、負の相関関係がある。

r が -1 に近い値であるほど、負の相関関係が強い。

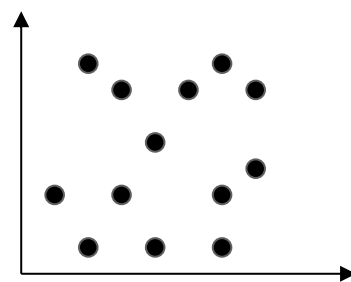
r が 0 に近い値であるほど、相関関係が弱い。



$r \doteq 0.73$



$r \doteq -0.81$



$r \doteq 0.14$

解答

$\bar{x} = \frac{1}{7}(10+8+4+6+9+5+7) = \frac{49}{7} = 7$, $\bar{y} = \frac{1}{7}(4+2+1+1+4+2+0) = \frac{14}{7} = 2$ から、次のような表を作る。

高校生	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
A	10	4	3	2	9	4	6
B	8	2	1	0	1	0	0
C	4	1	-3	-1	9	1	3
D	6	1	-1	-1	1	1	1
E	9	4	2	2	4	4	4
F	5	2	-2	0	4	0	0
G	7	0	0	-2	0	4	0
合計	49	14			28	14	14

したがって $r = \frac{14}{\sqrt{28} \sqrt{14}} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2} = 0.7$

このことから、 x と y の間には **強い正の相関関係がある** といえる。

1 1 仮説検定の考え方

同じくらいの速さで走る A さん、B さんがいる。
 この 2 人が競走をすると、いつも接戦になり
 勝敗も五分五分であった。
 あるとき A さんはシューズを変え、それから 15 回
 の競走は 12 勝 3 敗である。
 このことから、A さんはシューズによって速くなった
 と判断してよいか。仮説検定の考え方をを用い、
 基準となる確率を 0.05 とし考察せよ。
 ただし、コイン 15 枚を 1 度に投げ表が出た枚数
 を記録する実験を 1000 回くり返したところ、
 右の表のようになった。考察にはこの結果を用いよ。

表の枚数	度数
0	0
1	1
2	1
3	21
4	40
5	96
6	167
7	205
8	186
9	142
10	83
11	40
12	13
13	4
14	0
15	1
合計	1000

要 点

仮説検定の考え方

得られたデータをもとにある仮説を立て、それが正しいかどうかを判断する手法を **仮説検定** という。具体的には、得られたデータが予測していた結果と著しく異なる場合、それが偶然そうなったと考えるのではなく、予測の根拠がそもそも間違っているのではないかという考えのもとである仮説を立て、その仮説が妥当であるかどうかを判断する手法である。

仮説検定の手順

- ① 正しいかどうか判断したい主張に対して、その主張に反する仮説を立てる。
- ② 基準となる確率を定める。また、得られたデータが立てた仮説のもとでどの程度の確率で起こるかを求め、それらを比較する。
- ③ - (i) 基準となる確率より、得られたデータが起こる確率の方が小さくなる時、そもそも仮説が正しくなかったと判断する。
- ③ - (ii) 基準となる確率より、得られたデータが起こる確率の方が大きくなる時、仮説は正しいとも正しくないとも判断できない。

〈注意〉 (i)との対応から、「そもそも仮説が正しくなかったとは判断できない。」といってもよい。ただし、「仮説が正しい」とまではいえないことに注意する。

解答

正しいかどうか判断したい主張を、「Aさんはシューズによって速くなった。」とし、その主張に反する仮説を「シューズを変えた後もAさんとBさんの勝敗は五分五分である。」とする。

コイン投げの結果から、1000回中表が12枚以上出た度数は

$$13+4+0+1=18$$

よって、仮説のもとでAさんが12勝以上する確率は $\frac{18}{1000} = 0.018$ 程度であると考えられる。

これは0.05より小さいから、仮説「シューズを変えた後もAさんとBさんの勝敗は五分五分である。」は正しくなかったと考えられる。

したがって、Aさんはシューズによって速くなったと判断してよい。