1　データの分析

1.1　平均値，分散

問題

304 右の表は，あるクラスの生徒 10 人に対して行われた国語と英語の小テスト（各 10 点満点）の得点をまとめたものである。ただし，小テストの得点は整数値をとり，C > D である。また，表の数値はすべて正確な値であり，四捨五入されていない。

番号	国語	英語
生徒 1	9	9
生徒 2	10	9
生徒 3	4	8
生徒 4	7	6
生徒 5	10	8
生徒 6	5	C
生徒 7	5	8
生徒 8	7	9
生徒 9	6	D
生徒 10	7	7
平均値	A	8
分散	B	1

(1) 10 人の国語の得点について，平均値 A は［ア］点，分散 B の値は［イ］である。

(2) 10 人の英語の得点の平均値が 8 点，分散が 1 であることから，C と D の間には関係式

$$C + D = \boxed{ウ},$$

$$(C - 8)^2 + (D - 8)^2 = \boxed{エ}$$

が成り立つ。上の連立方程式と条件 C > D により，C，D の値は，それぞれ［オ］点，［カ］点であることがわかる。　（センター試験　改）

305 工場でつくられるある製品 25 個の長さを測ったら，次の数値（単位は mm）が得られた。

19.8	19.8	19.9	19.9	19.9	19.9	19.9	19.9	20.0	20.0
20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.1	20.1	20.1
20.1	20.1	20.2	20.2	20.3					

これらの製品の長さを x_i $(i = 1,\ 2,\ 3,\ \cdots,\ 25)$ とし，$y_i = 10(x_i - 20)$ とする。

(1) y_i $(i = 1,\ 2,\ 3,\ \cdots,\ 25)$ の平均 $\overline{y}$ と分散 v_y を求めよ。

(2) x_i $(i = 1,\ 2,\ 3,\ \cdots,\ 25)$ の平均 $\overline{x}$ と分散 v_x を求めよ。　（山梨大　改）

チェック・チェック

304 データ全体の分布の状態は，度数分布表やヒストグラムなどによって知ることができます。また，データの特徴を 1 つの値で表す代表値として平均値，中央値（メジアン），最頻値（モード）などがあります。

n 個のデータの値 x_1，x_2，…，x_n について

$$\text{平均値 } \overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

です。データの散らばりの度合いを表す量として分散があり，分散はデータの値 x_i と平均値 $\overline{x}$ の差（偏差）の平方の平均値として定義されます。

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

分散の正の平方根を標準偏差といいます。

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

305 変量 x，y の間に

$y = ax + b$ （a，b は定数）

という関係があるとき，x，y の平均値 $\overline{x}$，$\overline{y}$ と分散 ${s_x}^2$，${s_y}^2$ について次の等式が成り立ちます。

$$\overline{y} = a\overline{x} + b, \quad {s_y}^2 = a^2{s_x}^2$$

変量 x の仮平均を x_0 とすると，x と x_0 の差 $x - x_0$ の平均値 $\overline{u}$ は

$\overline{u} = \overline{x} - x_0$　したがって　$\overline{x} = \overline{u} + x_0$

となります。本問では，データの値の刻みが 0.1 なので

$$y = \frac{x - x_0}{0.1} = 10(x - x_0)$$

として y の値を整数にすることができ，さらに，x，y の平均値 $\overline{x}$，$\overline{y}$ の間に

$\overline{y} = 10(\overline{x} - x_0)$　したがって　$\overline{x} = \frac{1}{10}\overline{y} + x_0$

という関係が成り立ちます。

分散については，定義から次の等式が成り立ちます。

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

この式を利用すると計算がラクになります。

解答・解説

304 (1) 国語の得点の平均値 A は

$$A = \frac{1}{10}(9+10+4+7+10+5+5+7+6+7)$$
$$= \frac{70}{10} = \underline{\mathbf{7}}\ (\text{点})$$

であるから，国語の得点の分散 B は

$$B = \frac{1}{10}\{(9-7)^2+(10-7)^2+(4-7)^2+(7-7)^2+(10-7)^2$$
$$+(5-7)^2+(5-7)^2+(7-7)^2+(6-7)^2+(7-7)^2\}$$
$$= \frac{40}{10} = \underline{\mathbf{4}}$$

(2) 英語の得点の平均値が 8 点より

$$\frac{1}{10}(9+9+8+6+8+C+8+9+D+7) = 8$$
$$64+C+D = 8\times 10$$
$$\therefore\quad \underline{\mathbf{C+D=16}} \quad \cdots\cdots ①$$

英語の得点の分散が 1 より

$$\frac{1}{10}\{(9-8)^2+(9-8)^2+(8-8)^2+(6-8)^2+(8-8)^2$$
$$+(C-8)^2+(8-8)^2+(9-8)^2+(D-8)^2+(7-8)^2\} = 1$$
$$8+(C-8)^2+(D-8)^2 = 1\times 10$$
$$\therefore\quad \underline{\mathbf{(C-8)^2+(D-8)^2=2}} \quad \cdots\cdots ②$$

ここで，①より $D = 16-C$ であるから，これを②に代入して

$$(C-8)^2+\{(16-C)-8\}^2 = 2$$
$$2(C-8)^2 = 2$$
$$\therefore\quad C = 7,\ 9$$

よって

$$(C,\ D) = (7,\ 9),\ (9,\ 7)$$

$C > D$ より

$\underline{\mathbf{C=9}}$ (点)，　$\underline{\mathbf{D=7}}$ (点)

305 (1) y_i の度数分布表をつくると次のようになる。

y_i	-2	-1	0	1	2	3
度数	2	6	9	5	2	1

したがって，y_i の平均 $\overline{y}$ は

$$\begin{aligned}\overline{y} &= \frac{1}{25}\{(-2)\times 2+(-1)\times 6+0\times 9+1\times 5+2\times 2+3\times 1\}\\ &= \frac{2}{25}\\ &= \underline{\mathbf{0.08\ (mm)}}\end{aligned}$$

また，分散 v_y は

$$\begin{aligned}v_y &= ((y_i)^2 \text{ の平均}) - (\overline{y})^2 \quad \cdots\cdots ①\\ &= \frac{1}{25}\{(-2)^2\times 2+(-1)^2\times 6+0^2\times 9+1^2\times 5\\ &\qquad +2^2\times 2+3^2\times 1\}-0.08^2\\ &= \frac{36}{25}-0.08^2\\ &= 1.44-0.0064\\ &= \underline{\mathbf{1.4336}}\end{aligned}$$

【注意】 分散 v_y は y_i の偏差の 2 乗の平均を考えて

$$\begin{aligned}v_y = \frac{1}{25}\{&(-2-0.08)^2\times 2+(-1-0.08)^2\times 6+(0-0.08)^2\times 9\\ &+(1-0.08)^2\times 5+(2-0.08)^2\times 2+(3-0.08)^2\times 1\}\end{aligned}$$

から求めることもできるが，計算が煩雑なので ① を用いて計算するとよいだろう。

(2) $y_i = 10(x_i - 20)$ より

$$x_i = \frac{1}{10}y_i + 20$$

であるから，x_i の平均 $\overline{x}$ と分散 v_x は

$$\begin{aligned}\overline{x} &= \frac{1}{10}\overline{y}+20 = \frac{1}{10}\times 0.08+20\\ &= \underline{\mathbf{20.008\ (mm)}}\end{aligned}$$

$$\begin{aligned}v_x &= \left(\frac{1}{10}\right)^2 v_y = \frac{1}{100}\times 1.4336\\ &= \underline{\mathbf{0.014336}}\end{aligned}$$

である。

1.2　中央値・箱ひげ図

問題

306 ある町から 16 世帯を無作為に選んで所得を調べたところ，度数分布表は右のようになった。

所得（万円）	400	500	600	700	800	900	1000
度数（世帯数）	1	2	3	4	3	2	1

(1) 所得の平均値，中央値，最頻値および範囲を求めよ。

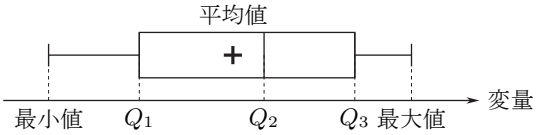
(2) 上の表で所得 1000 万円の世帯の所得が，実は 1000 万円でなく 9000 万円であったとする。このとき，所得の平均値，中央値，最頻値および範囲を求めよ。また，(1)，(2) の箱ひげ図を並べてかけ。

(3) (1)，(2) をもとに，平均値と中央値のどちらがどのような場合に資料の代表値として適切であるかを 150 字以内で述べよ。　（筑波大　改）

チェック・チェック

306 すべてのデータの値を小さい順に並べたとき，中央にくる値を**中央値**（メジアン）といいます。データの個数が奇数 $2n+1$ 個のときは $n+1$ 番目の値が中央値ですが，偶数 $2n$ 個のときは n 番目の値と $n+1$ 番目の値の平均値を中央値とします。また，データを度数分布表に整理したとき，度数が最も大きい階級の階級値を**最頻値**（モード）といい，データの最大値と最小値の差を**範囲**（レンジ）といいます。

データの分布を表す 1 つの方法として**箱ひげ図**があります。



すべてのデータの値を小さい方から並べたとき，全体を 4 等分する位置にある値を小さい方から第 1 四分位数，第 2 四分位数，第 3 四分位数といい，それぞれ Q_1，Q_2，Q_3 で表します。このとき，Q_2 は中央値です。Q_3-Q_1 を四分位範囲，$\dfrac{Q_3-Q_1}{2}$ を四分位偏差といいます。四分位偏差は Q で表します。

データ全体の中で，他のデータに比べて極端に値が大きかったり小さかったりするデータの値を**外れ値**といいます。平均値は外れ値の影響を大きく受けますが，中央値 Q_2 および Q_2-Q_1 と Q_3-Q_2 の平均である四分位偏差 Q は外れ値の影響を受けにくい値となっています。

解答・解説

306 (1) 所得の平均値は

$$\frac{1}{16}(400\times1+500\times2+600\times3+700\times4+800\times3+900\times2+1000\times1)$$
$$=\frac{1}{16}\times11200=\underline{\textbf{700（万円）}}$$

中央値は，所得を少ない順に並べたときの 8 番目の値と 9 番目の値の平均値であり

$$\frac{700+700}{2}=\underline{\textbf{700（万円）}}$$

最頻値は **700（万円）**

範囲は

$$1000-400=\underline{\textbf{600（万円）}}$$

(2) 所得の平均値は

$$\frac{1}{16}(11200-1000+9000)=\frac{1}{16}\times19200=\underline{\textbf{1200（万円）}}$$

中央値は，(1) と同じく **700（万円）**

最頻値は，(1) と同じく **700（万円）**

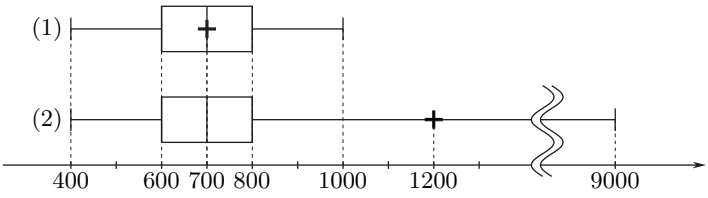
範囲は

$$9000-400=\underline{\textbf{8600（万円）}}$$

(1)，(2) の **5 数要約**をまとめると下表の通りである。

	最小値	第 1 四分位数 Q_1	第 2 四分位数 Q_2	第 3 四分位数 Q_3	最大値
(1)	400	600	700	800	1000
(2)	400	600	700	800	9000

これをもとに箱ひげ図をかくと **下図** となる。



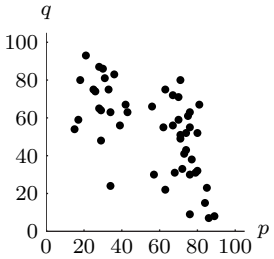
(3) 解答例は **以下の通り** である。

(2) のように，所得が極端に高い人がごく少数いる場合（外れ値がある場合），平均値を代表値とすることは必ずしも適切ではない。実際 (2) では，16 世帯中 15 世帯が平均値 1200 万円よりもはるかに低い所得となっている。このように，データの中に外れ値があるときは，中央値を代表値とする方が適切である。

1.3　相関係数

問題

307 (1) 変量 p と変量 q を観測した資料に対して，相関図（散布図）を作ったところ，右のようになった。ただし，相関図（散布図）中の点は，度数 1 を表す。2 つの変量 p と q の相関係数に最も近い値を，次の ⓪～⑥ のうちから一つ選べ。



⓪ -1.5　① -0.9　② -0.6　③ 0.0　④ 0.6　⑤ 0.9　⑥ 1.5

(2) 右の資料は 2 科目の小テストに関する 5 人の生徒の得点を記録したものである。2 科目の小テストの得点をそれぞれ変量 x，y とする。

生徒番号	1	2	3	4	5
x	3	4	5	4	4
y	7	9	10	8	6

(i) 変量 y を使って新しい変量 u を $u = ky$ $(k > 0)$ で定めると，変量 u の分散は x の分散と同じになる。このとき，k の値を求めよ。

(ii) 変量 x と変量 y の相関係数を r, 変量 x と変量 u の相関係数を r' とし，それぞれの 2 乗を r^2 と $(r')^2$ で表す。このとき，r^2 と $(r')^2$ の値をそれぞれ，小数第 2 位まで求めよ。　（センター試験　改）

チェック・チェック

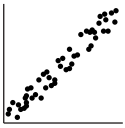
307 x の偏差と y の偏差の積の平均値を**共分散**といい，s_{xy} で表します。

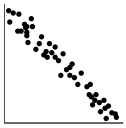
$$s_{xy} = \frac{1}{n}\{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots\cdots + (x_n - \overline{x})(y_n - \overline{y})\}$$

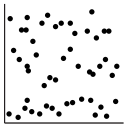
s_{xy} を x の標準偏差と y の標準偏差の積 $s_x s_y$ でわった値 r を**相関係数**といいます。

$$r = \frac{s_{xy}}{s_x s_y}, \quad -1 \leqq r \leqq 1$$

相関係数 r は，正の相関が強いほど値は 1 に近づき，負の相関が強いほど値は -1 に近づきます。


正の相関関係


負の相関関係


相関関係がない

解答・解説

307 (1) 相関係数の値は -1 と 1 の間にある。また，相関図の点は右下がりに分布しており，弱い負の相関関係がみられるから，選択肢として当てはまる値は $\underline{\boldsymbol{-0.6}\ (②)}$ である。

(2) (i) x，y の平均値はそれぞれ

$$\frac{3+4+5+4+4}{5}=4,\quad \frac{7+9+10+8+6}{5}=8$$

である。よって，x，y の分散をそれぞれ ${s_x}^2$，${s_y}^2$ とすると

$${s_x}^2=\frac{(3-4)^2+(4-4)^2+(5-4)^2+(4-4)^2+(4-4)^2}{5}=\frac{2}{5}$$

$${s_y}^2=\frac{(7-8)^2+(9-8)^2+(10-8)^2+(8-8)^2+(6-8)^2}{5}=2$$

ここで，u の分散を ${s_u}^2$ とすると，$u=ky$ より

$${s_u}^2=k^2{s_y}^2$$

が成り立つから，${s_x}^2={s_u}^2$ のとき

$$\frac{2}{5}=k^2\cdot 2 \qquad \therefore\quad \underline{\boldsymbol{k=\frac{\sqrt{5}}{5}}}\quad (\because\quad k>0)$$

(ii) x，y の共分散を s_{xy} とすると

$$s_{xy}=\frac{(-1)\cdot(-1)+0\cdot 1+1\cdot 2+0\cdot 0+0\cdot(-2)}{5}=\frac{3}{5}$$

であるから，x，y の相関係数 r に対して

$$r^2=\frac{{s_{xy}}^2}{{s_x}^2{s_y}^2}=\left(\frac{3}{5}\right)^2\div\left(\frac{2}{5}\cdot 2\right)=\frac{45}{100}=\underline{\boldsymbol{0.45}}$$

また，x，u の共分散を s_{xu} とすると，(i) の k を用いて

$$s_{xu}=ks_{xy}$$

であるから，x，u の相関係数 r' に対して

$$(r')^2=\frac{{s_{xu}}^2}{{s_x}^2{s_u}^2}=\frac{k^2{s_{xy}}^2}{{s_x}^2\cdot k^2{s_y}^2}=r^2=\underline{\boldsymbol{0.45}}$$