

第5章

データの分析

Check!

No.	難易度	1回目	2回目	No.	難易度	1回目	2回目
1	*			6	*		
2	*			7	**		
3	*			8	*		
4	*			9	**		
5	*						

例題

No.	難易度	1回目	2回目	No.	難易度	1回目	2回目
140	*			151	**		
141	*			152	**		
142	**			153	***		
143	**			154	***		
144	**						
145	*						
146	**						
147	**						
148	**						
149	***						
150	***						

練習

No.	難易度	1回目	2回目	No.	難易度	1回目	2回目
140	*			151	**		
141	*			152	**		
142	**			153	***		
143	**			154	***		
144	**						
145	*						
146	**						
147	**						
148	**						
149	***						
150	***						

Step Up

No.	難易度	1回目	2回目	No.	難易度	1回目	2回目
1	**			11	***		
2	**			12	**		
3	**			13	*		
4	**			14	***		
5	**			15	***		
6	**						
7	***						
8	***						
9	***						
10	***						

まとめ

1

データの整理と分析

1. 度数分析

- 変量……………ある特性を数量的に表すもの

データ……………変量の測定値を集めたもの

度数分布表……データの値の区間を設定し、その区間に入るデータの値の個数を数えてまとめたもの

階級……………度数分布表で設定される区間

階級の幅……………区間の幅

階級値……………各階級の両端の平均値

度数……………各階級に含まれるデータの値の個数
- ◀ 個々の値をデータの値という。

◀ 階級は一般に等間隔で作成するが、両端の階級などには大きな階級幅を設定する場合もある。

2. 相対度数・累積度数

- 相対度数……各階級の度数の全体に占める割合

年度が違う調査の比較など、複数の度数分布表を比べる場合、データ全体の個数の相違が問題となるが、相対度数分布表であれば比較しやすい。

相対度数分布表では、各階級の相対度数の総和は1となる。

累積度数……………最初の階級からその階級までの度数を合計したもの

累積相対度数……最初の階級からその階級までの相対度数を合計したもの
- ◀ (相対度数)
$$= \frac{(\text{その階級の度数})}{(\text{度数の合計})}$$

【注】 四捨五入して各階級の相対度数を求めると、誤差の関係で、資料によっては相対度数の合計がちょうど1.00にならない場合があるが、全体を1として考えているという意味で、相対度数の合計欄には、1.00と書けばよい。また、合計がちょうど1.00となるように各階級の相対度数を調節する必要はないが、円グラフに表したりするなど、特にそれが必要な場合には、相対度数の最も大きな階級で調節することが多い。

(例) 30人のクラスのテストの点数について、階級幅を10点としたときの相対度数分布表を考える。

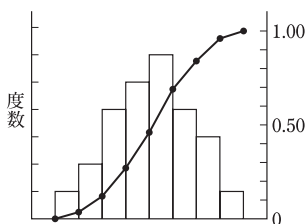
階級	度数	相対度数	累積度数	累積相対度数
30 以上 40 未満	2	0.07	2	0.07
40 以上 50 未満	3	0.10	5	0.17
50 以上 60 未満	5	0.17	10	0.34
60 以上 70 未満	8	0.27	18	0.61
70 以上 80 未満	9	0.30	27	0.91
80 以上 90 未満	3	0.10	30	1.00
合計	30	1.00		

◀ 相対度数の合計は、
$$0.07 + 0.10 + 0.17 + 0.27 + 0.30 + 0.10 = 1.01$$

となるが、合計欄には1.00と書けばよい。

3. ヒストグラム

度数分布表をもとに、横軸に階級の値、縦軸に度数を取り、右図のように各階級の度数を柱状のグラフで表したものをヒストグラムという。



▶ ヒストグラムでは、グラフの各長方形は、隣接させてかく。

また、階級の右端を横軸、累積相対度数を縦軸にとり、右図のように各階級の累積相対度数を線分でつないでできる折れ線グラフを累積相対度数折れ線グラフという。

4. 代表値

代表値……………データ全体の傾向を、適当な1つの数値で表したもの

代表値としてよく使われるものに、平均値、中央値、最頻値がある。

平均値……………データの値の総和を総度数で割ったもの

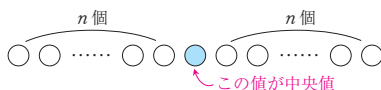
$$(\text{平均値}) = \frac{(\text{データの値の総和})}{(\text{総度数})}$$

中央値 (メジアン)…データの値を大きさの順に並べたとき、中央にくる値

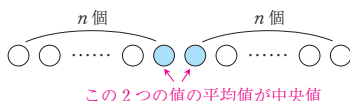
最頻値 (モード)……最も度数が多いデータの値

【注】 中央値はデータの値の個数によって、次の場合がある。

(1) データの値の個数が奇数 ($2n+1$) のとき



(2) データの値の個数が偶数 ($2n$) のとき
中央にある2つの値の平均値を中央値とする。



▶ 変数 x の平均値を \bar{x} と表すことがある。

▶ 仮に設定した基準の値をもとに平均値を求める方法もある。この仮の基準値を、仮平均という。(p.285 参照)

5. 四分位数と箱ひげ図

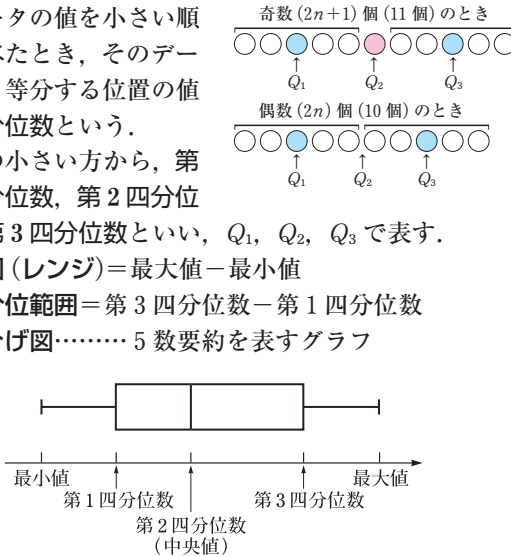
データの値を小さい順に並べたとき、そのデータを4等分する位置の値を四分位数という。

値の小さい方から、第1四分位数、第2四分位数、第3四分位数といい、 Q_1 、 Q_2 、 Q_3 で表す。

範囲(レンジ)=最大値-最小値

四分位範囲=第3四分位数-第1四分位数

箱ひげ図……5数要約を表すグラフ



外れ値…他の値から極端にかけ離れた値。外れ値の目安は、 Q_1 から小さい方(または Q_3 から大きい方)へ四分位範囲の1.5倍以上離れていること

最小値、第1四分位数、第2四分位数、第3四分位数、最大値の5つの数値をまとめて5数要約という。

箱の中に「+」をかき入れ、平均値を表すこともある。

測定ミスなど原因がわかっているものは異常値と呼び、外れ値と区別することもある。

6. 分散と標準偏差

n 個のデータの値(x_1, x_2, \dots, x_n)があり、その平均値を \bar{x} とする。

$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$ を、それぞれ x_1, x_2, \dots, x_n の偏差という。

偏差の2乗の平均値を分散といい、 s^2 で表す。

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

また、 $s^2 = \overline{x^2} - (\bar{x})^2 = (\overline{x^2} \text{の平均値}) - (\bar{x} \text{の平均値})^2$ でも計算できる。

分散の正の平方根を標準偏差といい、 s で表す。

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

平均値

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

偏差を2乗したものを偏差平方という。

Standard deviation (標準偏差)

7. 変量の変換

変量 x から a , b を定数として, $y=ax+b$ によって新しい変量 y が得られるとき, y の平均値を \bar{y} , 分散を s_y^2 , 標準偏差を s_y とすると,

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

8. 散布図

2 個の変量からなる 2 次元データを平面にプロットした図のことを散布図 (相関図) という。

2 個の変量の間に,

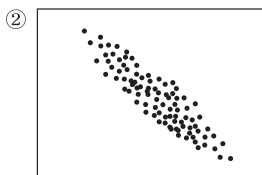
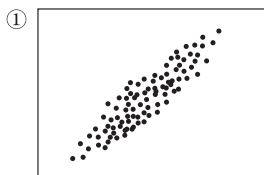
正の相関がある……一方が増えると他方も直線的に
増える傾向がみられる。

データを表す点は全体的に
右上がりになる。→ ①

負の相関がある……一方が増えると他方が直線的に
減る傾向がみられる。

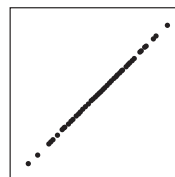
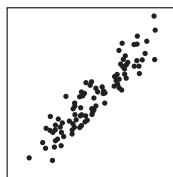
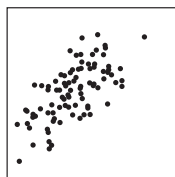
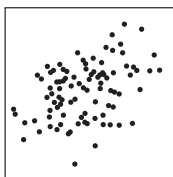
データを表す点は全体的に
右下がりになる。→ ②

直線的な増減でなければ相関はない。

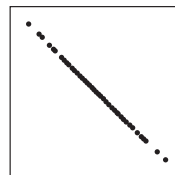
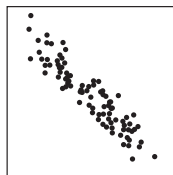
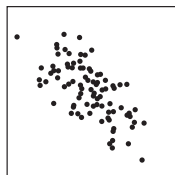
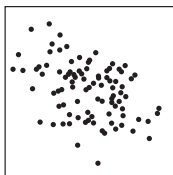


①でも②でもないとき, 相関はないという。

散布図において, 直線的な傾向がはっきり表れている
ものほど, 相関関係は強いという。



正の相関が強い



負の相関が強い

9. 相関係数

2 個の変量からなる n 個の 2 次元データを,
 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
とする.

x_1, x_2, \dots, x_n の平均値を \bar{x} , 標準偏差を s_x ,
 y_1, y_2, \dots, y_n の平均値を \bar{y} , 標準偏差を s_y とする.

$(x_1 - \bar{x})(y_1 - \bar{y}), (x_2 - \bar{x})(y_2 - \bar{y}), \dots, (x_n - \bar{x})(y_n - \bar{y})$
のことを偏差積という.

この平均値を変量 x と変量 y の共分散といい, s_{xy} で表す.

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

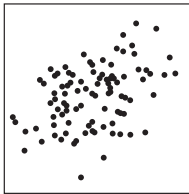
共分散が正のとき → 変量 x と変量 y の間には
正の相関がある.

共分散が負のとき → 変量 x と変量 y の間には
負の相関がある.

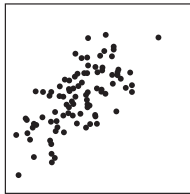
相関関係の方向と強弱を表す指標として, 相関係数 r

という量を考え, $r = \frac{s_{xy}}{s_x s_y}$ で定義する.

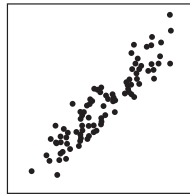
相関係数 r については, 一般に, $-1 \leq r \leq 1$ が成り立つ.
相関係数 r の値が 1 に近いほど, 正の相関が強くなる.



$r=0.3$

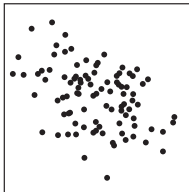


$r=0.6$

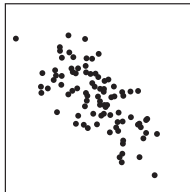


$r=0.9$

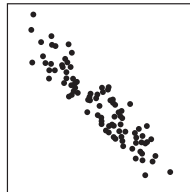
相関係数 r の値が -1 に近いほど, 負の相関が強くなる.



$r=-0.3$



$r=-0.6$



$r=-0.9$

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$s_x = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

$$s_y = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_n - \bar{y})^2}{n}}$$

x や y の単位を変更しても, r の値は変わらない.

10. 相関と因果

一方が原因でもう一方が結果となるような関係を因果関係という。

AとBに相関がみられるとき、AとBの関係について次のような場合が考えられる。

- ① 因果關係 $A \rightarrow B$ ② 因果關係 $B \rightarrow A$

- ③ 共通の要因 $C \begin{cases} \nearrow A \\ \searrow B \end{cases}$ ④ その他

相関があるからといって、必ずしも因果関係があるとはいえない。

→は、原因から結果の
流れを表す。

11. 仮説検定の考え方

仮説検定…あるデータが与えられたとき、仮説を立てて、それが妥当かを判定する統計の手法

仮説検定の流れ

仮説 A を立てる



仮説Aを否定する仮説Bを考える



仮説Bを前提として、与えられたデータの割合を考える



割合が極めて小さい→仮説Bを否定し、
仮説Aが妥当と判断

割合が小さいわけでない→結論を保留

仮説Aを対立仮説，仮説Bを帰無仮説という。
(数学Bで学習)

割合が小さいかどうかの基準はあらかじめ決めておく。

割合が小さいわけでないときは、仮説Bを否定することができないが、仮説Aを否定することもできない。

12. 統計的探求プロセス

統計的探求プロセス…様々な課題を統計を通して解決する一連のプロセス

統計的探求プロセスの流れ

→問題の発見 (problem)…問題の把握, 問題設定



調査の計画 (plan)・・・データの想定、収集計画



データの収集 (data) ... データの収集, 表への整理



分析 (analysis) … グラフの読み取り, 傾向の把握



—結論 (conclusion)…結論付け、振り返り

PPDAC サイクルともいう。

結論からさらなる課題
が出る場合などは、も
う一度PPDACサイ
クルを行うとよい。

Check!

*

- 1 8人の生徒に10点満点のテストを行ったところ、次のような結果になった。

6, 2, 5, 9, 3, 7, 5, 4(点)

- (1) 階級幅を2点として度数分布表を作り、ヒストグラムをかけ。ただし、最初の階級を0点以上2点未満とする。
- (2) 最も度数の多い階級の階級値を求めよ。

*

- 2 ある小学校の通学団の学年構成を調べたところ、6年生…1人、3年生…1人、2年生…1人、1年生…2人であった。次の表を完成させよ。

階級(学年)	度数(人)	累積度数(人)	相対度数	累積相対度数
1以上3未満				
3以上5未満				
5以上6以下				
合計				

*

- 3 次の表は、あるクラスの生徒8名の100m走の計測結果である。

出席番号	1	2	3	4	5	6	7	8
記録(秒)	13.2	19.5	12.7	11.2	13.6	13.8		12.3

出席番号7の生徒のデータの値が消失しているが、平均値が13.8秒であることがわかっている。

- (1) 出席番号7の生徒のデータの値を求めよ。
- (2) このデータの中央値を求めよ。

*

- 4 次のデータの最頻値と四分位数を求め、箱ひげ図をかけ。

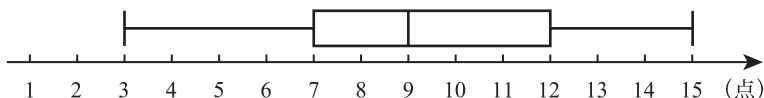
- (1) 6, 5, 4, 7, 8, 6, 6, 8, 5

(2)

階級値	2	6	10	14	計
度数	2	3	2	1	8

*

- 5 次の箱ひげ図は、36人の小テストの結果を表している。



点数の高い順にして、上から10番目の人のテストの結果は何点以上何点以下だといえるか。

*

6

あるレストランでランチサービスを始めた。最近5日間の注文数を調べたところ、次の通りであった。

90, 110, 140, 70, 90 (食)

- (1) 注文数の平均値を求めよ。(2) 注文数の分散と標準偏差を求めよ。

**

7

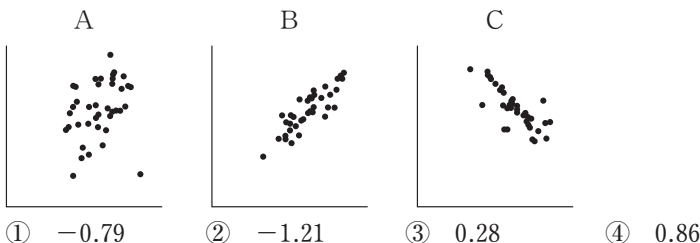
変数 x の n 個のデータ x_1, x_2, \dots, x_n がある。

x の平均値 \bar{x} が6であるとき、変数 $y = \frac{1}{4}x + 20$ の平均値 \bar{y} を求めよ。

*

8

次の散布図A～Cのそれぞれに対し、最も近いと思われる相関係数 r の値を、下の①～④から選べ。



**

9

2つの変数 x と y が次の表で与えられている。

x	4	8	0	2	1
y	2	5	1	4	3

- (1) x を横軸に、 y を縦軸にとって散布図をかけ。
 (2) x と y の平均値をそれぞれ求めよ。
 (3) 次の表を完成させよ。

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
4	2					
8	5					
0	1					
2	4					
1	3					

- (4) x と y の相関係数 r を求めよ。

▶▶ 解答編 p. 225

第5章

- 1 (1) 略 (2) 5点 2 略 3 (1) 14.1秒 (2) 13.4秒 4 (1) 最頻値6, 第1四分位数5, 第2四分位数6, 第3四分位数7.5 (2) 最頻値6, 第1四分位数4, 第2四分位数6, 第3四分位数10 5 9点以上12点以下 6 (1) 100食 (2) 分散560, 標準偏差 $4\sqrt{35}$ 食 7 $\bar{y} = 21.5$ 8 A ③ B ④ C ① 9 (1) 略
 (2) x の平均値3, y の平均値3 (3) 略 (4) $r = \frac{7}{10}$

Think

例題

140

相対度数分布表

太郎さんの作った模型飛行機はどれくらいの距離を飛ぶのか、実際に30回の測定をして記録したものが、以下の数値である。

49, 67, 64, 50, 24, 59, 46, 67, 56, 30, 64, 17, 48, 71, 69,
60, 78, 68, 66, 56, 58, 73, 61, 50, 37, 66, 53, 25, 56, 46 (m)

(1) 表の空いているところに太郎さんの記録を入れ、表を完成させよ。

階級 (m)	度数 (個)	相対度数	度数 (個)	相対度数
0 以上 10 未満	0	0	0	0
10 以上 20 未満	1	0.03	3	0.06
20 以上 30 未満			2	0.04
30 以上 40 未満			9	0.18
40 以上 50 未満			14	0.28
50 以上 60 未満			7	0.14
60 以上 70 未満			11	0.22
70 以上 80 未満			4	0.08
合計	30	1.00	50	1.00

(2) 上の表の右側の度数分布表と相対度数分布表は、先生が作った模型飛行機の飛距離の記録である。次の文章は正しいといえるか。

- ① 70 m 以上の記録が先生の方が多いため、先生の飛行機の方がよく飛ぶといえる。
- ② 最も相対度数の大きい階級に着目すると、太郎さんの飛行機の方がよく飛ぶといえる。

考え方

相対度数は、全体に対する割合を考えたものなので、度数の異なるデータを比較するときに利用するとよい。

解答

(1)

階級 (m)	度数 (個)	相対度数	度数 (個)	相対度数
0 以上 10 未満	0	0	0	0
10 以上 20 未満	1	0.03	3	0.06
20 以上 30 未満	2	0.07	2	0.04
30 以上 40 未満	2	0.07	9	0.10
40 以上 50 未満	4	0.13	14	0.28
50 以上 60 未満	8	0.27	7	0.14
60 以上 70 未満	10	0.33	11	0.22
70 以上 80 未満	3	0.10	4	0.08
合計	30	1.00	50	1.00

- (2) 70 m 以上の記録の相対度数を比べると、太郎さんは 0.10 で先生は 0.08 なので先生の飛行機の方がよく飛ぶとはいえない。

一方、相対度数の最も大きい階級は、太郎さんは 60 m 以上 70 m 未満で、先生は 40 m 以上 50 m 未満なので、太郎さんの方がよく飛ぶといえる。

よって、①は正しいとはいえない。②は正しいといえる。

【注】例題 140 の太郎さんの記録について、累積相対度数を調べると次のようになる。

階級 (m)	度数 (個)	相対度数	累積度数	累積相対度数
0 以上 10 未満	0	0	0	0
10 以上 20 未満	1	0.03	1	0.03
20 以上 30 未満	2	0.07	3	0.10
30 以上 40 未満	2	0.07	5	0.17
40 以上 50 未満	4	0.13	9	0.30
50 以上 60 未満	8	0.27	17	0.57
60 以上 70 未満	10	0.33	27	0.90
70 以上 80 未満	3	0.10	30	1.00
合計	30	1.00		

累積度数や累積相対度数では、階級ごとにではなく、どこからどこまでの階級についてなど、範囲を限定した度数や割合を調べることができる。たとえば、例題 140 の太郎さんの記録の場合、60 m 未満のときの累積相対度数が 0.57 になるので、60 m 未満の場合と 60 m 以上の場合がおおよそ半分程度ずつになるとみることができる。

練習 140

*

花子さんに目分量で 10 cm の線分を 40 本かいてもらい、実際の長さを計測した。それを記録したものが以下の数値である。次の表を完成させ、花子さんがかいた線分の傾向として、①、②の文章が正しいといえるか。

9.3, 8.8, 9.7, 8.8, 9.3, 9.1, 8.9, 8.3, 8.6, 9.4,
9.9, 10.9, 8.7, 10.6, 10.3, 8.4, 9.6, 11.2, 10.8, 10.0,
8.6, 9.4, 10.4, 8.3, 8.8, 10.0, 8.6, 9.4, 9.0, 9.0,
8.8, 9.8, 8.7, 9.6, 9.2, 8.2, 9.5, 8.9, 8.5, 9.9 (cm)

階級 (cm)	度数 (本)	相対度数	累積度数	累積相対度数
8.0 以上 8.5 未満	4	0.10	4	0.10
8.5 以上 9.0 未満				
9.0 以上 9.5 未満				
9.5 以上 10.0 未満				
10.0 以上 10.5 未満				
10.5 以上 11.0 未満				
11.0 以上 11.5 未満				
11.5 以上 12.0 未満				
合計	40	1.00		

- ① 度数が最も多い階級は、9.5 cm 以上 10.0 cm 未満である。

- ② 半分程度は 10 cm の誤差 5 mm の長さである。

例題 141 データの傾向

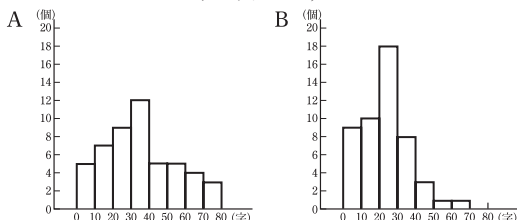
2人の小説家 A, B の書く文章の特徴を調べるために「文の長さ」に注目した。それぞれが書いた小説から 50 の文を無作為に選び、その文字数を記録し度数分布表にまとめたものが、以下の数値である。

A	階級(字)	度数(個)	B	階級(字)	度数(個)
	0 以上 10 未満	5		0 以上 10 未満	9
	10 以上 20 未満	7		10 以上 20 未満	10
	20 以上 30 未満	9		20 以上 30 未満	18
	30 以上 40 未満	12		30 以上 40 未満	8
	40 以上 50 未満	5		40 以上 50 未満	3
	50 以上 60 未満	5		50 以上 60 未満	1
	60 以上 70 未満	4		60 以上 70 未満	1
	70 以上 80 未満	3		70 以上 80 未満	0
	合計	50		合計	50

A, B の書く文章を比較したとき、どのようなことがいえるか。

考え方 データの値の個数が多いとき、度数分布表に整理すると、その傾向がわかりやすくなる。度数分布表の階級幅は、大きすぎず、小さすぎず、全体の傾向がよく表されるように選ばよい。

解答 ヒストグラムをかくと次のようになる。



よって、A の書く文に比べて、B の書く文の方が短い傾向がある。とくに B は、20 字以上 30 字未満の文を書くことが多い。

▶ 例題 141 の解は度数分布表やヒストグラムから読み取れることであればよいので、たとえば、「A が書く文は 30 字以上 40 字未満の文が一番多いのに対して、B の書く文は 20 字以上 30 字未満の文が一番多い」などでもよい。

練習 141

A, B の 2 人に目分量で 10 cm の線分を 25 本ずつかいてもらい、実際の長さを計測した。それを記録したものが、以下の数値である。

*

A 8.7, 9.1, 10.3, 8.6, 9.4, B 10.0, 10.7, 11.0, 10.1, 9.8,
 9.7, 8.4, 9.0, 8.8, 9.9, 10.5, 9.4, 10.4, 11.5, 11.3,
 8.9, 10.7, 8.8, 9.5, 8.7, 9.7, 9.3, 10.5, 9.8, 10.2,
 9.8, 9.3, 8.5, 8.9, 10.4, 10.0, 9.9, 10.5, 10.4, 9.8,
 8.6, 10.2, 9.7, 9.2, 9.0 (cm) 9.8, 9.3, 10.3, 9.6, 10.0 (cm)

A, B のかく線分を比較したとき、どのようなことがいえるか。

例題 142 代表値と度数分布表(1)

次のデータは、ある商品の20日間の販売数を1日ごとに並べたものである。

10 9 11 8 14 11 12 9 10 13
12 11 9 12 8 13 10 8 7 11 (個)

- (1) このデータについて、下の表を完成させ、仮平均を10個として、販売数の平均値を求めよ。

販売数(個)	7	8	9	10	11	12	13	14
日数(日)								

- (2) さらに5日を加えた25日間の販売数の平均値を、11個以上とするためには、残り5日間で何個以上売り上げる必要があるか。

考え方 (1) n 個のデータの値 $x_1, x_2, x_3, \dots, x_n$ の平均値 \bar{x} を仮平均 x_0 を用いて表すと、
 $\bar{x} = x_0 + \frac{1}{n}\{(x_1 - x_0) + (x_2 - x_0) + \dots + (x_n - x_0)\}$ となる。

解答

(1)

販売数(個)	7	8	9	10	11	12	13	14
日数(日)	1	3	3	3	4	3	2	1

仮平均を10個とすると、20日間のデータの平均値は、

$$10 + \frac{1}{20}\{(-3) \times 1 + (-2) \times 3 + (-1) \times 3 + 0 \times 3 + 1 \times 4 + 2 \times 3 + 3 \times 2 + 4 \times 1\}$$

$$= 10 + \frac{8}{20} = 10 + 0.4 = 10.4 \text{ (個)}$$

- (2) 20日間の販売数は、 $10.4 \times 20 = 208$ (個)

1日に11個売れたときの25日間の販売数は、

$$11 \times 25 = 275 \text{ (個)}$$

したがって、残り5日間で x 個売り上げるとして、

25日間の販売数の平均値が11個以上となるとき、

$$208 + x \geq 275$$

よって、 $x \geq 67$ より、**67個以上**

(平均値) \times (総度数)
 $=$ (総和)

練習
142

**

次のデータは、ある商品の20日間の生産数を1日ごとに並べたものである。

20 23 21 20 21 20 19 20 17 22
17 22 18 19 21 20 18 19 19 18 (個)

- (1) このデータについて、下の表を完成させ、仮平均を20個として、販売数の平均値を求めよ。

生産数(個)	17	18	19	20	21	22	23
日数(日)							

- (2) さらに10日を加えた30日間の生産数の平均値を、20個以上とするためには、残り10日間で何個以上生産する必要があるか。

例題 143 代表値と度数分布表(2)

次の表は、生徒 40 人の試験の得点 (0 以上の整数) の累積度数をまとめたもので、各生徒の得点は明らかではない。このとき、次の問いに答えよ。

得点 (点)	90 以上	80 以上	70 以上	60 以上	50 以上	40 以上	30 以上	20 以上
度数 (人)	0	3	12	26	32	36	39	40

- (1) 80 点以上 90 点未満を 1 つの階級として、各階級値に対する度数分布表を作成せよ。
- (2) (1) で作成した度数分布表における平均値を求めよ。
- (3) 生徒 40 人の実際の得点の平均値の最大値と最小値を求めよ。

考え方 (3) データの平均値 \bar{x} の最大値と最小値は、
最大 (小) 値：各データの値が各階級の最大 (小) 値をとったときの平均値

解答 (1)

階級値 (点)	85	75	65	55	45	35	25
度数 (人)	3	9	14	6	4	3	1

▶ 階級値は各階級の両端の平均値である。

- (2) 平均値は、

$$\begin{aligned} & \frac{1}{40}(85 \times 3 + 75 \times 9 + 65 \times 14 + 55 \times 6 + 45 \times 4 + 35 \times 3 + 25 \times 1) \\ &= \frac{2480}{40} = 62 \text{ (点)} \end{aligned}$$

- (別解) 仮平均を最頻値 65 点とすると、平均値は、

$$\begin{aligned} & 65 + \frac{1}{40}\{20 \times 3 + 10 \times 9 + 0 \times 14 + (-10) \times 6 + (-20) \times 4 \\ & \quad + (-30) \times 3 + (-40) \times 1\} \\ &= 65 - \frac{120}{40} = 65 - 3 = 62 \text{ (点)} \end{aligned}$$

- (3) 各データの値が各階級の最大値をとるとき、すなわち、各データの値が各階級の階級値より 4 点だけ大きい値となるとき、平均値は最大となるから、平均値の**最大値**は、 $62 + 4 = 66$ (点)

同様に、各データの値が各階級の階級値より 5 点だけ小さい値となるとき、平均値は最小となるから、平均値の**最小値**は、 $62 - 5 = 57$ (点)

注 仮平均は最頻値や中央値に近い数にとることが多い。また、平均値を実際のデータから求めたときと、度数分布表から求めたときとは、必ずしも結果は一致しない。

練習 143

**

次の表は、生徒 100 人の試験の得点 (0 以上の整数) の累積度数をまとめたもので、各生徒の得点は明らかではない。このとき、次の問いに答えよ。

得点 (点)	90 以上	80 以上	70 以上	60 以上	50 以上	40 以上	30 以上	20 以上	10 以上
度数 (人)	0	7	20	38	63	85	93	98	100

- (1) 80 点以上 90 点未満を 1 つの階級として、各階級値に対する度数分布表を作成せよ。
- (2) (1) で作成した度数分布表における平均値を求めよ。
- (3) 生徒 100 人の実際の得点の平均値のとり得る値の範囲を求めよ。▶ p.309 ①

例題 144 代表値と四分位数

次の表は、生徒 13 人の A 班と生徒 12 人の B 班に 10 点満点の試験を行った結果である。

得点(点)	0	1	2	3	4	5	6	7	8	9	10	平均値
A 班(人)	0	0	0	2	1	1	4	2	2	1	0	x
B 班(人)	0	1	1	1	0	y	z	1	2	1	0	5.5

- (1) 表中の x , y , z の値を求めよ。
- (2) それぞれの班のデータについて、中央値、四分位範囲を求めよ。
- (3) A 班と B 班をあわせた 25 人の平均値を求めよ。

考え方 (2) A 班の人数は奇数、B 班の人数は偶数であることに注意して中央値を求める。

解答 (1) x は A 班の平均値であるから、

$$x = \frac{1}{13}(3 \times 2 + 4 \times 1 + 5 \times 1 + 6 \times 4 + 7 \times 2 + 8 \times 2 + 9 \times 1) = \frac{78}{13} = 6$$

B 班の人数と平均値より、

$$1 + 1 + 1 + y + z + 1 + 2 + 1 = 12 \quad \cdots \cdots \textcircled{1}$$

$$\frac{1}{12}(1 \times 1 + 2 \times 1 + 3 \times 1 + 5 \times y + 6 \times z + 7 \times 1 + 8 \times 2 + 9 \times 1) = 5.5 \quad \cdots \cdots \textcircled{2}$$

①, ②より, $y + z = 5$, $5y + 6z = 28$ であるから, $y = 2$, $z = 3$

- (2) A 班のデータの値を小さい順に並べると、

3 3 4 5 6 6 6 6 7 7 8 8 9

A 班の中央値は、6 点

A 班の第 1 四分位数は、 $\frac{4+5}{2} = 4.5$ (点)、第 3 四分位数は、 $\frac{7+8}{2} = 7.5$ (点)

であるから、A 班の四分位範囲は、 $7.5 - 4.5 = 3$ (点)

B 班のデータの値を小さい順に並べると、

1 2 3 5 5 6 6 6 7 8 8 9

B 班の中央値は、 $\frac{6+6}{2} = 6$ (点)

B 班の第 1 四分位数は、 $\frac{3+5}{2} = 4$ (点)、第 3 四分位数は、 $\frac{7+8}{2} = 7.5$ (点)

であるから、B 班の四分位範囲は、 $7.5 - 4 = 3.5$ (点)

- (3) A 班と B 班をあわせた平均値は、 $\frac{1}{25}(6.0 \times 13 + 5.5 \times 12) = 5.76$ (点)

第5章

練習 144

**

次の表は、生徒 39 人を A 班と B 班に分け、10 点満点の試験を行った結果である。

得点(点)	0	1	2	3	4	5	6	7	8	9	10	平均値
A 班(人)	0	0	1	1	3	3	4	2	2	2	1	x
B 班(人)	0	1	1	2	1	y	z	5	4	2	0	6

- (1) 表中の x , y , z の値を求めよ。また、39 人全員の点数の平均値を求めよ。
- (2) それぞれの班のデータについて、中央値、四分位範囲を求めよ。

例題 145 四分位数と箱ひげ図(1)

次の表は、生徒 45 人に 10 点満点のテストを行った結果である。

得点(点)	0	1	2	3	4	5	6	7	8	9	10
人数(人)	0	0	1	4	2	6	12	14	3	2	1

このデータについて、四分位範囲を求めよ。また、このデータの箱ひげ図をかけ。

考え方 箱ひげ図をかくために必要な情報は 5 数要約である。四分位範囲はどの部分を示すか考える。

解答 第 2 四分位数は中央値で、23 番目の値だから、6 点
第 1 四分位数は、11 番目の値 5 と 12 番目の値 5 の平均

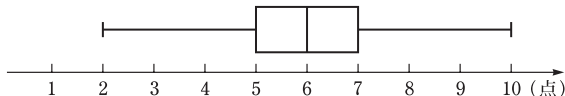
値だから、 $\frac{5+5}{2}=5$ (点)

第 3 四分位数は、34 番目の値 7 と 35 番目の値 7 の平均

値だから、 $\frac{7+7}{2}=7$ (点)

したがって、四分位範囲は、 $7-5=2$ (点)

また、最小値は 2 点、最大値は 10 点であるから、箱ひげ図は、次のようになる。



$x_1, \dots, x_{11}, x_{12}, \dots, x_{22}$
11 個 11 個

$x_{24}, \dots, x_{34}, x_{35}, \dots, x_{45}$
11 個 11 個

(四分位範囲)
=(第 3 四分位数)
-(第 1 四分位数)

Focus

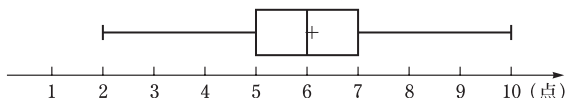
データの個数が偶数が奇数かに注意する

注 例題 145 の生徒 45 人のテストの平均点を求めると次のようになる。

$$\frac{1}{45}(2 \times 1 + 3 \times 4 + 4 \times 2 + 5 \times 6 + 6 \times 12 + 7 \times 14 + 8 \times 3 + 9 \times 2 + 10 \times 1)$$

$$= 6.08 \dots \approx 6.1 \text{ (点)}$$

これを箱ひげ図にかき入れる場合は、次のように「+」を入れる。



練習 145

*

次の表は、ある野球チームの最近 30 試合の得点の結果である。

得点(点)	0	1	2	3	4	5	6	7	8
試合数(試合)	3	6	4	2	8	4	2	0	1

このデータについて、四分位範囲を求めよ。また、このデータの箱ひげ図をかけ。ただし、箱ひげ図には平均値もかき入れるものとする。

→ p. 309 ② ③

例題 146

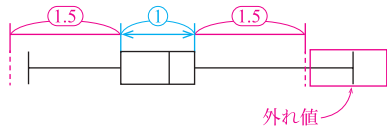
四分位数と箱ひげ図(2)

次のデータは、家庭菜園で育てたミニトマトの収穫数を年ごとに18年間記録したものである。このデータの箱ひげ図をかけ。ただし、外れ値がある場合は、外れ値を示して箱ひげ図をかけ。

12, 22, 32, 29, 26, 40, 27, 33, 36,
34, 40, 42, 32, 42, 50, 72, 56, 55 (個)

考え方 外れ値は、第1四分位数から小さい方、もしくは第3四分位数から大きい方へ四分位範囲の1.5倍以上離れていることが目安となる。
まずは、与えられたデータを小さい順に並べ、5数要約と四分位範囲を調べる。

外れ値の目安



解答 与えられたデータを小さい順に並べると、次のようになる。

12, 22, 26, 27, 29, 32, 32, 33, 34,
36, 40, 40, 42, 42, 50, 55, 56, 72
よって、最大値は72個、最小値は12個

第1四分位数は、29個

第2四分位数は、 $\frac{34+36}{2}=35$ (個)

第3四分位数は、42個

四分位範囲は、 $42-29=13$ (個)

ここで、 $13 \times 1.5 = 19.5$ より、外れ値の目安は、

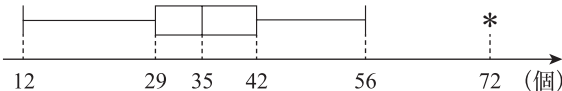
$29 - 19.5 = 9.5$ より、9.5 以下か、

$42 + 19.5 = 61.5$ より、61.5 以上

のデータである。

したがって、72個が外れ値となる。

よって、箱ひげ図は次のようになる。



第2四分位数は中央値

四分位範囲の1.5倍

第1四分位数より小さいか第3四分位数より大きいかで判断する。

外れ値を除いて最大の値が最大値となる。その他の5数要約はそのまま示す。

第5章

練習 146

**

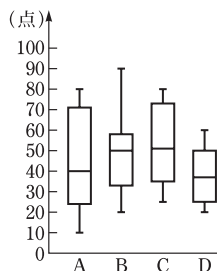
次のデータは、16日間、毎日の血圧を測定した値である。このデータの箱ひげ図をかけ。

ただし、外れ値がある場合は外れ値を示し、箱ひげ図をかけ。

50, 55, 58, 64, 65, 66, 66, 67,
68, 69, 70, 72, 73, 78, 80, 89 (mmHg)

例題 147 箱ひげ図

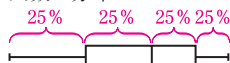
右の図は、生徒数がいずれも 40 人の A 組, B 組, C 組, D 組に、100 点満点の同じテストを行った結果を箱ひげ図に表したものである。



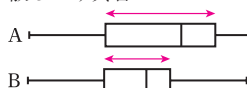
- (1) 50 点以上の生徒が 20 人以上いる組はどれか。
- (2) 上位 10 人の散らばりが最も大きい組はどれか。
- (3) 65 点をとった生徒が上位から 15 番目、30 点をとった生徒が上位から 25 番目であった組はどれか。
- (4) 全体の散らばりが最も小さい組はどれか。

考え方 箱ひげ図からは次のことが読み取れる。

(ア) 人数の分布



(イ) 散らばり具合



全長は同じだが、Aの方が箱の長さは長いので散らばりも大きい。

解答 各組の生徒数は 40 人であるから、中央値は点数の低い方から 20 番目と 21 番目の得点の平均値である。

第 1 四分位数は点数の低い方から 10 番目と 11 番目の得点の平均値、第 3 四分位数は点数の高い方から 10 番目と 11 番目の得点の平均値である。

- (1) 生徒 20 人は、各組の生徒数 40 人の 50% に相当する。

つまり、中央値が 50 点以上の組であるから、**B 組と C 組**である。

- (2) 上位 10 人の散らばりが最も大きい、つまり、箱ひげ図の上側のひげの長さが最も長い組は、**B 組**である。

- (3) 上位から 15 番目は中央値 Q_2 と第 3 四分位数 Q_3 の間に位置し、上位から 25 番目は第 1 四分位数 Q_1 と中央値 Q_2 の間に位置する。

箱ひげ図が、 $Q_1 < 30 < Q_2 < 65 < Q_3$ となっている組は、**A 組**である。

- (4) 全体の散らばりは範囲 (箱ひげ図の全長: 箱とひげを合わせた部分) で決まる。

全体の散らばりが最も小さい、つまり、全長が最も短いのは、**D 組**である。

▶ 点数の低い方から 26 番目

▶ 点数の低い方から 16 番目

▶ 範囲の大小から散らばりを考える。

練習 147

例題 147 のデータにおいて、次の問いに答えよ。

- (1) 60 点以上の生徒が 10 人以下の組はどれか。
- (2) 上位 10 人が 70 点以上で、その散らばりが最も小さい組はどれか。
- (3) 全体の散らばりが最も大きい組はどれか。

→ p. 309 [2]

Think

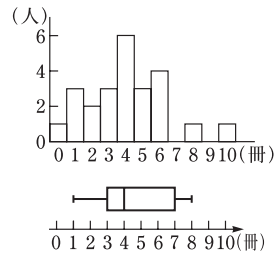
例題

148

ヒストグラムと箱ひげ図

右のヒストグラムは、生徒 24 人の P 組の 1 か月の読書冊数を表したものである。

- (1) P 組について、箱ひげ図を作れ。
- (2) 右の図は、生徒 24 人の Q 組の 1 か月の読書冊数の箱ひげ図である。Q 組と P 組の箱ひげ図から読み取れることとして次の①～③は正しいといえるか。



- ① Q 組より P 組の方が上位 25% の読書冊数の散らばりが大きい。
- ② 四分位範囲に含まれる生徒数は、P 組と比べて Q 組の方が多い。
- ③ 1 か月で読書冊数が 7 冊以上の生徒が Q 組には 6 人以上いる。

考え方

2 つの変量の傾向を箱ひげ図から読み取るときは、箱ひげ図を並べて表記するとよい。

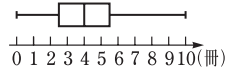
解答

- (1) 中央値は、 $\frac{4+4}{2}=4$ (冊)

第 1 四分位数は、 $\frac{2+3}{2}=2.5$ (冊)

第 3 四分位数は、 $\frac{5+6}{2}=5.5$ (冊)

よって、箱ひげ図は右ようになる。



- (2) ① 箱ひげ図の右側のひげの長さは、Q 組より P 組の方が長く、上位 25% は P 組の方が散らばりが大きいといえるから、正しい。
- ② 四分位範囲に含まれる生徒の割合は全体の 50% であり、P 組、Q 組ともに生徒数が 24 人より、四分位範囲に含まれる生徒数はどちらも同じであるから、正しくない。
- ③ Q 組の第 3 四分位数は 7 冊であり、読書冊数が第 3 四分位数以上の生徒は Q 組全体の 25% である。また、Q 組の全生徒は 24 人で、その 25% は、 $24 \times 0.25 = 6$ (人) であるから、1 か月で読書冊数が 7 冊以上の生徒が Q 組には 6 人以上いる。したがって、正しい。

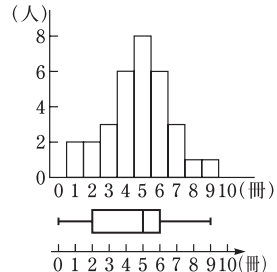
練習

148

**

右のヒストグラムは、生徒 32 人の A 組の 1 か月の読書冊数を表したものである。

- (1) A 組について、箱ひげ図を作れ。
- (2) 右の図は、生徒 32 人の B 組の 1 か月の読書冊数の箱ひげ図である。A 組と B 組の箱ひげ図から読み取れることとして次の①～③は正しいといえるか。



- ① A 組と B 組の合計 64 人の生徒の 25% は 6 冊以上の本を読んでいる。
- ② A 組の下位 25% の人数は、B 組の下位 25% の人数よりも多い。
- ③ A 組と B 組の最頻値はどちらも 5 冊である。

→ p. 310 ④

例題

149

分散と標準偏差

- (1) 変量 x の n 個のデータの値 x_1, x_2, \dots, x_n がある. x の平均値を \bar{x} , x^2 の平均値を $\overline{x^2}$ とすると, x の分散 s^2 は, $s^2 = \overline{x^2} - (\bar{x})^2$ と表せることを証明せよ.
- (2) 次の表は, A 組と B 組で同じテストを行った結果であり, この表を使って A 組と B 組の平均値を求めると, ともに 5.3 点であった.

得点 (点)	0	1	2	3	4	5	6	7	8	9	10	合計
A 組 (人)	0	0	0	2	4	6	4	2	2	0	0	20
B 組 (人)	0	2	2	2	2	2	2	3	3	1	1	20

この表から, A 組と B 組の標準偏差をそれぞれ求めよ. また, A 組と B 組の得点の散らばりを比較するとどのようなことがいえるか.

考え方

- (1) 分散の定義 $s^2 = \frac{1}{n} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2\}$ を利用して, 式を変形する.
- (2) 分散の正の平方根が標準偏差である.
変量 x の分散を s^2 とすると,

$$s^2 = \overline{x^2} - (\bar{x})^2 = \frac{1}{20} \{ \text{(各生徒の得点の平方の和)} - \text{(平均値)}^2 \}$$

解答

- (1) 分散の値 s^2 は,

$$\begin{aligned}
 s^2 &= \frac{1}{n} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2\} \\
 &= \frac{1}{n} \{(x_1^2 + x_2^2 + \dots + x_n^2) \\
 &\quad - 2\bar{x}(x_1 + x_2 + \dots + x_n) + n(\bar{x})^2\} \\
 &= \frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2) \\
 &\quad - 2\bar{x} \cdot \frac{1}{n} (x_1 + x_2 + \dots + x_n) + \frac{n}{n} (\bar{x})^2 \quad \dots\dots \textcircled{1}
 \end{aligned}$$

ここで,

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

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

であるから, ①に代入して,

$$\begin{aligned}
 s^2 &= \overline{x^2} - 2\bar{x} \cdot \bar{x} + (\bar{x})^2 \\
 &= \overline{x^2} - 2(\bar{x})^2 + (\bar{x})^2 \\
 &= \overline{x^2} - (\bar{x})^2
 \end{aligned}$$

よって, $s^2 = \overline{x^2} - (\bar{x})^2$ と表せることが示された.

偏差平方の平均値が分散である.

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

(2) (1)より, A組の分散は,

$$\begin{aligned} & \frac{1}{20}(3^2 \times 2 + 4^2 \times 4 + 5^2 \times 6 + 6^2 \times 4 + 7^2 \times 2 + 8^2 \times 2) \\ & \quad - (5.3)^2 \\ &= \frac{602}{20} - \left(\frac{53}{10}\right)^2 = \frac{3010 - 2809}{100} = \frac{201}{100} \end{aligned}$$

よって, A組の標準偏差は,

$$\sqrt{\frac{201}{100}} = \frac{\sqrt{201}}{10} \text{ (点)}$$

B組の分散は,

$$\begin{aligned} & \frac{1}{20}(1^2 \times 2 + 2^2 \times 2 + 3^2 \times 2 + 4^2 \times 2 + 5^2 \times 2 + 6^2 \times 2 \\ & \quad + 7^2 \times 3 + 8^2 \times 3 + 9^2 \times 1 + 10^2 \times 1) - (5.3)^2 \\ &= \frac{702}{20} - \left(\frac{53}{10}\right)^2 = \frac{3510 - 2809}{100} = \frac{701}{100} \end{aligned}$$

よって, B組の標準偏差は,

$$\sqrt{\frac{701}{100}} = \frac{\sqrt{701}}{10} \text{ (点)}$$

標準偏差はB組の方が大きいので, B組はA組よりデータの散らばりが大きいといえる.

Focus

変量 x の n 個のデータの値 x_1, x_2, \dots, x_n について,
 x の平均値を \bar{x} とするとき,
 x の分散 s^2 は,

$$\textcircled{1} \quad s^2 = \frac{1}{n} \{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \}$$

$$\textcircled{2} \quad s^2 = \overline{x^2} - (\bar{x})^2$$

【注】(2)のA組の分散について, 定義 (Focus の①の式) に基づいて計算すると,

$$\begin{aligned} & \frac{1}{20} \{ (3-5.3)^2 \times 2 + (4-5.3)^2 \times 4 + (5-5.3)^2 \times 6 \\ & \quad + (6-5.3)^2 \times 4 + (7-5.3)^2 \times 2 + (8-5.3)^2 \times 2 \} \end{aligned}$$

となり, ()² の部分を計算するのが少し面倒であるので, Focus の②の式を用いるのが有効である.

練習

149

次の表は, P組とQ組で同じテストを行った結果である.

得点 (点)	0	1	2	3	4	5	6	7	8	9	10	合計
P組 (人)	0	2	2	3	6	7	5	3	1	1	0	30
Q組 (人)	1	6	2	3	2	5	6	1	3	1	0	30

この表から, P組とQ組の標準偏差をそれぞれ求めよ. また, P組とQ組の得点の散らばりを比較するとどのようなことがいえるか.

→ p. 310 ⑤ ~ ⑧

Think

例題

150

変量の変換

- (1) p, q を定数とする. 変量 x に対し, $x = pu + q$, すなわち $u = \frac{x-q}{p}$ で定められる変量 u の平均値を \bar{u} , u の標準偏差を s_u とすると, 変量 x の平均値 \bar{x} は $\bar{x} = p\bar{u} + q$, 標準偏差 s_x は $s_x = |p|s_u$ であることを証明せよ.
- (2) 次の変量 x のデータについて, $u = \frac{x-100}{5}$ において得られる新しい変量 u の標準偏差 s_u を計算し, 変量 x の標準偏差 s_x を求めよ.

120, 115, 95, 105, 90, 125, 85, 80, 100, 105

考え方

変量 x のデータが $x_1, x_2, x_3, \dots, x_n$ であるとき, 平均値を \bar{x} とする.
変量 y が $y = ax$ の場合と, $y = x + b$ の場合について考えてみる.

- (i)
- $y = ax$
- の場合 (
- a
- は定数)

変量 y のデータは, $ax_1, ax_2, ax_3, \dots, ax_n$ となるので,

$$\begin{aligned}\text{平均値 } \bar{y} \text{ は, } \bar{y} &= \frac{1}{n}(ax_1 + ax_2 + ax_3 + \dots + ax_n) \\ &= a \cdot \frac{1}{n}(x_1 + x_2 + x_3 + \dots + x_n) = a\bar{x}\end{aligned}$$

- (ii)
- $y = x + b$
- の場合 (
- b
- は定数)

変量 y のデータは, $x_1 + b, x_2 + b, x_3 + b, \dots, x_n + b$ となるので,

$$\begin{aligned}\text{平均値 } \bar{y} \text{ は, } \bar{y} &= \frac{1}{n}(x_1 + b + x_2 + b + x_3 + b + \dots + x_n + b) \\ &= \frac{1}{n}\{(x_1 + x_2 + x_3 + \dots + x_n) + nb\} \\ &= \frac{1}{n}(x_1 + x_2 + x_3 + \dots + x_n) + b = \bar{x} + b\end{aligned}$$

このことから, $y = ax + b$ (a, b は定数) の場合の y の平均値 \bar{y} も $\bar{y} = a\bar{x} + b$ となるのがわかる. 標準偏差や分散についても同様に考えればよい.

解答

- (1)
- $k = 1, 2, \dots, n$
- とすると, 変量
- x
- のデータの値

$$\text{が } x_k \text{ のとき, } u_k = \frac{x_k - q}{p} \text{ とおくと, } x_k = pu_k + q$$

このとき,

$$\begin{aligned}\bar{x} &= \frac{1}{n}(x_1 + x_2 + \dots + x_n) \\ &= \frac{1}{n}\{(pu_1 + q) + (pu_2 + q) + \dots + (pu_n + q)\} \\ &= p \cdot \frac{1}{n}(u_1 + u_2 + \dots + u_n) + \frac{1}{n} \cdot nq \\ &= p\bar{u} + q\end{aligned}$$

$$\begin{aligned}&\frac{1}{n}(u_1 + u_2 + \dots + u_n) \\ &= \bar{u}\end{aligned}$$

また、変量 x , u の分散をそれぞれ s_x^2 , s_u^2 とすると、

$$\begin{aligned}
 s_x^2 &= \frac{1}{n} \{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \} \\
 &= \frac{1}{n} \{ [(pu_1 + q) - (p\bar{u} + q)]^2 \\
 &\quad + [(pu_2 + q) - (p\bar{u} + q)]^2 \\
 &\quad + \cdots + [(pu_n + q) - (p\bar{u} + q)]^2 \} \\
 &= \frac{1}{n} \{ (pu_1 - p\bar{u})^2 + (pu_2 - p\bar{u})^2 \\
 &\quad + \cdots + (pu_n - p\bar{u})^2 \} \\
 &= p^2 \cdot \frac{1}{n} \{ (u_1 - \bar{u})^2 + (u_2 - \bar{u})^2 \\
 &\quad + \cdots + (u_n - \bar{u})^2 \} \\
 &= p^2 s_u^2
 \end{aligned}$$

$\leftarrow \frac{1}{n} \{ (u_1 - \bar{u})^2 + (u_2 - \bar{u})^2 + \cdots + (u_n - \bar{u})^2 \} = s_u^2$

$$s_x \geq 0, s_u \geq 0 \text{ より, } s_x = \sqrt{p^2 s_u^2} = |p| s_u$$

(2) $u = \frac{x-100}{5}$ より、右のような表を作る。

これより、 u の平均値 \bar{u} は、

$$\begin{aligned}
 \bar{u} &= \frac{1}{10} \{ 4 + 3 + (-1) + 1 + (-2) + 5 \\
 &\quad + (-3) + (-4) + 0 + 1 \} \\
 &= \frac{4}{10} = \frac{2}{5}
 \end{aligned}$$

u^2 の平均値 \bar{u}^2 は、

$$\begin{aligned}
 \bar{u}^2 &= \frac{1}{10} (16 + 9 + 1 + 1 + 4 \\
 &\quad + 25 + 9 + 16 + 0 + 1) \\
 &= \frac{82}{10} = \frac{41}{5}
 \end{aligned}$$

u の分散 s_u^2 は、

$$s_u^2 = \bar{u}^2 - (\bar{u})^2 = \frac{41}{5} - \left(\frac{2}{5}\right)^2 = \frac{201}{25}$$

u の標準偏差 s_u は、

$$s_u = \sqrt{s_u^2} = \sqrt{\frac{201}{25}} = \frac{\sqrt{201}}{5}$$

以上より、 x の標準偏差 s_x は、

$$s_x = |5| s_u = 5 \times \frac{\sqrt{201}}{5} = \sqrt{201}$$

x	$x-100$	u	u^2
120	20	4	16
115	15	3	9
95	-5	-1	1
105	5	1	1
90	-10	-2	4
125	25	5	25
85	-15	-3	9
80	-20	-4	16
100	0	0	0
105	5	1	1

\leftarrow (1)の結果を利用

練習
150

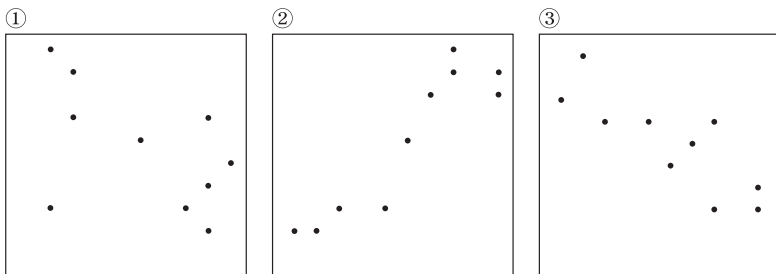
次の変量 x のデータについて、 $u = \frac{x-60}{5}$ において得られる新しい変量 u の標

準偏差 s_u を計算し、変量 x の標準偏差 s_x を求めよ。

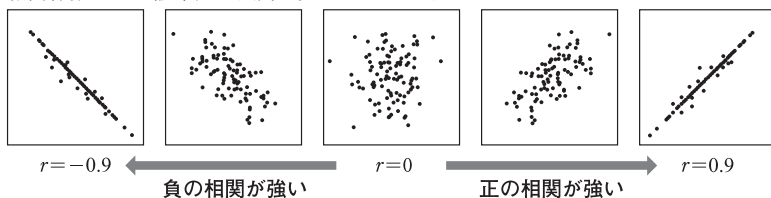
40, 40, 55, 50, 60, 65, 75, 75, 80, 85

例題 151 散布図と相関係数(1)

生徒 10 人に 2 種類のテスト A, B を行ったところ、得点の分散はどちらも $\frac{46}{5}$ で、A と B の得点の共分散は $\frac{42}{5}$ であった。このとき、A と B の得点の相関係数を求めよ。また、A と B の得点の散布図としてふさわしいものを次の①～③から選べ。



考え方 相関係数 r から散布図の傾向を見ることができる。



解答 A の得点の分散が $\frac{46}{5}$ より、標準偏差は $\sqrt{\frac{46}{5}}$ 点
 B の得点の分散が $\frac{46}{5}$ より、標準偏差は $\sqrt{\frac{46}{5}}$ 点
 A と B の得点の共分散は $\frac{42}{5}$ より、求める相関係数は、

$$\frac{\frac{42}{5}}{\sqrt{\frac{46}{5}} \times \sqrt{\frac{46}{5}}} = \frac{42}{46} = \frac{21}{23}$$

また、相関係数は正で、正の相関があるから、散布図は②

相関係数

$$= \frac{A, B \text{ の共分散}}{S_A \cdot S_B}$$

S_A : A の標準偏差

S_B : B の標準偏差

S_{xy} : A, B の共分散

正の相関があるとき、散布図上の点は全体的に右上がりになる。

練習 151

**

例題 151 において、さらに生徒 10 人がテスト C を行ったところ、C の得点の分散が $\frac{23}{5}$ で、A と C の得点の共分散が $-\frac{27}{5}$ であった。A と C の得点の相関係数を求めよ。また、A と C の得点の散布図としてふさわしいものを、例題 151 の①～③から選べ。

Column コラム

「散布図と度数分布表」

散布図は2種類のデータの全体の傾向を見るのにより方法であるが、2種類の数値的データの関係を調べる場合、2次元の度数分布表を用いる方法もある。

(問題) ある高校の生徒10名に、懸垂とボール投げのテストを行った。

出席番号	1	2	3	4	5	6	7	8	9	10
懸垂(回)	4	3	4	7	10	2	8	9	6	5
ボール投げ(m)	27	25	25	28	30	24	27	29	28	26

懸垂の結果を横軸に、ボール投げの結果を縦軸にとって散布図をかけ。

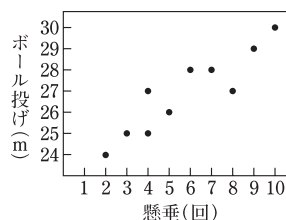
このデータをもとに散布図をかくと、右の図の

ようになり、右の散布図から、

「ボール投げの距離が遠い人ほど、

懸垂の回数も多い」

傾向が見られる。



また、度数分布表を作ると、次のようになる。

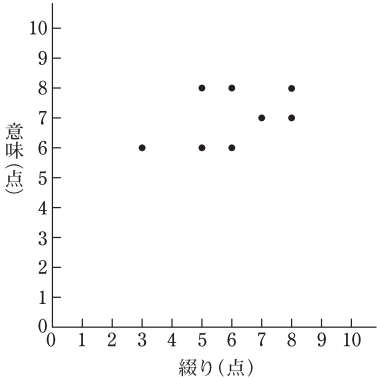
	5回未満	5回以上 10回未満	10回以上	合計
27 m 未満	3	1	0	4
27 m 以上 30 m 未満	1	4	0	5
30 m 以上	0	0	1	1
合計	4	5	1	10

この表では、横の並び(行)にボール投げ、縦の並び(列)に懸垂のデータがまとめられている。たとえば、ボール投げが27 m 以上 30 m 未満で、懸垂が5回以上 10回未満の度数は4となっている。このように、2次元の度数分布表で表すと、たとえば、ボール投げの記録が同じ階級の生徒の中で、懸垂の回数の割合を求めることができる。また、次のように行の和を100%で表すと、ボール投げにおけるある階級についての懸垂の分布を調べることができる。

	5回未満	5回以上 10回未満	10回以上	合計
27 m 未満	75.0%	25.0%	0.0%	100%
27 m 以上 30 m 未満	20.0%	80.0%	0.0%	100%
30 m 以上	0.0%	0.0%	100.0%	100%
合計	40.0%	50.0%	10.0%	100%

例題 152 散布図と相関係数(2)

右の図は、8人の生徒に行った英単語の綴りと意味を問うテスト(ともに10点満点)の得点の散布図で、綴りの得点を横軸に、意味の得点を縦軸にとったものである。



- (1) 次の表は、8人の生徒の出席番号、綴りと意味の得点をまとめた表である。空欄をうめ、表を完成させよ。

出席番号	1	2	3	4	5	6	7	8	平均値
綴り(点)	3			6	5	8	6	5	
意味(点)	6		8	6	6	7	8		

- (2) この8人の綴りと意味の得点の標準偏差がそれぞれ $\frac{\sqrt{10}}{2}$ 点、 $\frac{\sqrt{3}}{2}$ 点で、共分散が $\frac{5}{8}$ である。綴りと意味の相関係数 r を求めよ。
- (3) 意味の採点にミスはなかったが、綴りの採点にミスがあり、出席番号1と5の生徒の綴りの点数がともに4点に変更された。変更後の綴りと意味の相関係数 R を求めよ。

考え方

- (3) 変更後の綴りの標準偏差と、綴りと意味の共分散を求める。
その際、綴りの平均値は、変更前と変更後で同じである点に着目する。

解答

(1)

出席番号	1	2	3	4	5	6	7	8	平均値
綴り(点)	3	7	8	6	5	8	6	5	6
意味(点)	6	7	8	6	6	7	8	8	7

(2) $r = \frac{5}{8} \div \left(\frac{\sqrt{10}}{2} \times \frac{\sqrt{3}}{2} \right) = \frac{5}{2\sqrt{30}} = \frac{\sqrt{30}}{12}$

- (3) 変更前と変更後の綴りの点数を表にすると、次のようになる。

出席番号	1	2	3	4	5	6	7	8	平均値
綴り(変更前)	3	7	8	6	5	8	6	5	6
綴り(変更後)	4	7	8	6	4	8	6	5	6

変更後の綴りの平均値は、変更前と変わらない。
変更後の綴りの得点の分散は、

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

$$\frac{1}{8} \left[\left(\frac{\sqrt{10}}{2} \right)^2 \times 8 - \{(3-6)^2 + (5-6)^2\} + \{(4-6)^2 + (4-6)^2\} \right]$$

$$= \frac{18}{8} = \frac{9}{4}$$

したがって、変更後の綴りの標準偏差は、

$$\sqrt{\frac{9}{4}} = \frac{3}{2} \text{ (点)}$$

変更後の綴りと意味の得点の共分散は、

$$\begin{aligned} & \frac{1}{8} \left[\frac{5}{8} \times 8 - \{(3-6)(6-7) + (5-6)(6-7)\} \right. \\ & \quad \left. + \{(4-6)(6-7) + (4-6)(6-7)\} \right] \\ & = \frac{5}{8} \end{aligned}$$

よって、変更後の綴りと意味の相関係数 R は、

$$R = \frac{5}{8} \div \left(\frac{3}{2} \times \frac{\sqrt{3}}{2} \right) = \frac{5}{6\sqrt{3}} = \frac{5\sqrt{3}}{18}$$

分散や共分散を最初から計算し直してもよいが、ここでは変更前と変更後で平均値が同じであることを利用して、計算量を減らしている。

$$\begin{aligned} & \frac{1}{8} \{ (\text{変更前の綴りの分散}) \times 8 \\ & \quad - (\text{変更箇所の変更前の綴りの偏差平方和}) \\ & \quad + (\text{変更箇所の変更後の綴りの偏差平方和}) \} \end{aligned}$$

$$\begin{aligned} & \frac{1}{8} \{ (\text{変更前の共分散}) \times 8 \\ & \quad - (\text{変更箇所の変更前の偏差積和}) \\ & \quad + (\text{変更箇所の変更後の偏差積和}) \} \end{aligned}$$

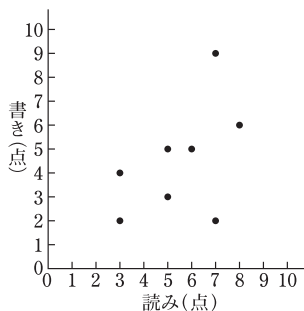
練習 152

**

右の図は、8人の生徒に行った漢字の読みと書きを問うテスト（ともに10点満点）の得点の散布図で、読みの得点を横軸に、書きの得点を縦軸にとったものである。

- (1) 次の表は、8人の生徒の出席番号、読みと書きの得点をまとめた表である。空欄をうめ、表を完成させよ。

出席番号	1	2	3	4	5	6	7	8	平均値
読み(点)	3	8	3			7		5	
書き(点)	2		4	5	3	2	9	5	



- (2) この8人の読みと書きの得点の標準偏差がそれぞれ $\sqrt{3}$ 点、 $\frac{\sqrt{19}}{2}$ 点で、共分散が $\frac{15}{8}$ である。読みと書きの相関係数 r を求めよ。
- (3) 読みの採点にはミスがなかったが、書きの採点にミスがあり、出席番号1, 2, 3, 4の生徒の書きの点数がそれぞれ1点ずつ加算された。変更後の読みと書きの相関係数 R を求めよ。

Column コラム

「相関と因果」

【問題1】

「『ごみを捨てないようにしましょう』という看板やポスターが多い地域ほど、ごみが路上に落ちていることが多い」

というデータがあったとする。このとき、

「看板やポスターが多い」ことが原因で「路上のごみが多い」
という関係があるといつてよいだろうか。

【問題2】

「全国のサラリーマンについて、体脂肪率と所得の相関をとると、正の相関がみられた」

というデータがあったとする。このとき、

「体脂肪が上がる」ことが原因で「所得も多くなる」
という関係があるといつてよいだろうか。

問題1では、「路上のごみが多い」から「看板やポスターが多い」と考えるのが自然で、問題2では、たとえば、「年齢」という別の要因があり、年齢が上がると、体脂肪率が上がる傾向や、所得が多くなる傾向があると考えるのが自然であろう。

一般に、AとBに相関がみられるとき、AとBの関係については、次のような場合が考えられる。

- ① 因果関係 $A \rightarrow B$ (Aが原因でBである)
- ② 因果関係 $B \rightarrow A$ (Bが原因でAである)
- ③ 共通の要因 $C \begin{cases} \nearrow A \\ \searrow B \end{cases}$ (Cが原因でAであり、Bである)
- ④ その他

問題1では②、問題2では③にあたると考えられる。つまり、問題1も2も①の因果関係があるとはいいいきれない。

このように、相関がみられるからといって、そこに因果関係があるとはいえない場合があるので、データを分析してその傾向などから要因を探りたいときには注意が必要である。

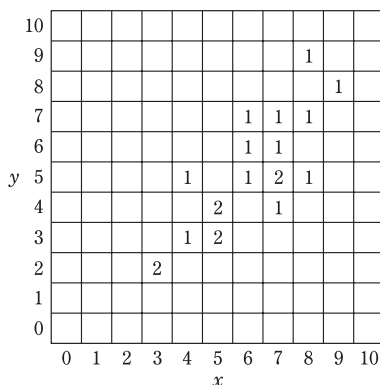
Think

例題

153

総合問題

右の図は、生徒 20 人に行った関数と図形のテスト（ともに 10 点満点）の結果をまとめたものである。関数の得点 x を横軸に、図形の得点 y を縦軸にとっている。図の中の数値は x, y の値の組に対応する人数を表している。たとえば、関数の得点が 7 点で図形の得点が 5 点である生徒の人数は 2 人である。



- (1) 各生徒の得点について、 $x+y$ の最大値と、 $|x-y|$ の最大値を求めよ。
- (2) 図をもとに、次の表を完成させよ。また、各テストの得点の平均値を求めよ。

得点(点)	0	1	2	3	4	5	6	7	8	9	10
関数(人)											
図形(人)											

- (3) (2)の表を使って各テストの標準偏差を求めると、関数は $\sqrt{2.8}$ 点、図形は $\sqrt{3.6}$ 点で、関数と図形の得点の共分散は 2.55 であった。関数と図形の得点の相関係数 r の値を四捨五入して小数第 2 位まで求めよ。ただし、 $\sqrt{7}=2.646$ とする。
- (4) 右の表は、別の 5 人の生徒 A, B, C, D, E に同じ問題のテストを行った結果である。5 人の関数と図形の得点の平均値は、それぞれ 20 人の得点の平均値と同じであった。20 人にこの 5 人を加えた合計 25 人の生徒に関する関数と図形の得点の相関係数 R の値を小数第 2 位まで求めよ。

5 人の生徒	A	B	C	D	E
関数の得点	7	4	6	9	4
図形の得点	5	4	5	6	5

- (5) これらのテストの結果について、次の①～③は正しいといえるか。
- ① 生徒 25 人の得点について、関数と図形の平均値からの散らばり具合は同じである。
- ② 生徒 20 人の関数と図形の得点の正の相関はやや強いが、A～E の 5 人が加わると正の相関は少し弱まる。
- ③ 生徒 25 人の図形の得点が一様に 1 点上がれば、25 人の関数と図形の得点の相関係数 R の値はより大きくなる。

解答

- (1) $x+y$ が最大となる生徒の x, y の値の組は, $(x, y)=(8, 9), (9, 8)$ より, $x+y$ の最大値は 17 である.

$|x-y|$ が最大となる生徒の x, y の値の組は, $(x, y)=(7, 4), (8, 5)$ より, $|x-y|$ の最大値は 3 である.

(2)

得点 (点)	0	1	2	3	4	5	6	7	8	9	10
関数 (人)	0	0	0	2	2	4	3	5	3	1	0
図形 (人)	0	0	2	3	3	5	2	3	1	1	0

関数の得点の平均値は,

$$\frac{1}{20}(3 \times 2 + 4 \times 2 + 5 \times 4 + 6 \times 3 + 7 \times 5 + 8 \times 3 + 9 \times 1) = \frac{120}{20} = 6 \text{ (点)}$$

図形の得点の平均値は,

$$\frac{1}{20}(2 \times 2 + 3 \times 3 + 4 \times 3 + 5 \times 5 + 6 \times 2 + 7 \times 3 + 8 \times 1 + 9 \times 1) = \frac{100}{20} = 5 \text{ (点)}$$

$$\begin{aligned} (3) \quad r &= \frac{2.55}{\sqrt{2.8} \sqrt{3.6}} = \frac{25.5}{\sqrt{28} \sqrt{36}} = \frac{25.5}{12\sqrt{7}} = \frac{25.5 \times \sqrt{7}}{84} = \frac{25.5 \times 2.646}{84} \\ &= \frac{67.473}{84} \div 0.80 \end{aligned}$$

- (4) 5 人の関数と図形の得点の平均値がそれぞれ 20 人の平均値と同じであることから, 生徒 25 人の関数と図形の得点の平均値も, これらと同じになる.

20 人の関数の得点の偏差平方の和は,

$$(\sqrt{2.8})^2 \times 20 = 56$$

また, 5 人の関数の得点の偏差平方の和は,

$$(7-6)^2 + (4-6)^2 + (6-6)^2 + (9-6)^2 + (4-6)^2 = 18$$

であるから, 生徒 25 人の関数の得点の分散 s_x^2 は,

$$s_x^2 = \frac{1}{25}(56 + 18) = \frac{74}{25}$$

20 人の図形の得点の偏差平方の和は,

$$(\sqrt{3.6})^2 \times 20 = 72$$

また, 5 人の図形の得点の偏差平方の和は,

$$(5-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (5-5)^2 = 2$$

であるから, 生徒 25 人の図形の得点の分散 s_y^2 は,

$$s_y^2 = \frac{1}{25}(72 + 2) = \frac{74}{25}$$

20 人の得点の偏差積の和は,

$$2.55 \times 20 = 51$$

また, 5 人の得点の偏差積の和は,

$$(7-6)(5-5) + (4-6)(4-5) + (6-6)(5-5) + (9-6)(6-5) + (4-6)(5-5) = 5$$

したがって, 25 人の関数と図形の得点の共分散 s_{xy} は,

$$s_{xy} = \frac{1}{25}(51 + 5) = \frac{56}{25}$$

よって, 生徒 25 人の関数と図形の得点の相関係数

R は,

$$R = \frac{s_{xy}}{s_x s_y} = \frac{56}{25} \div \left(\sqrt{\frac{74}{25}} \times \sqrt{\frac{74}{25}} \right) = \frac{56}{74} = \frac{28}{37} \div 0.76$$

▶ (20 人の得点の
偏差平方の和)
= (20 人の分散) \times 20

▶ $\frac{1}{25}$ {(20 人の得点の
偏差平方の和)
+ (5 人の得点の
偏差平方の和)}

▶ $\frac{1}{25}$ {(20 人の得点の
偏差積の和)
+ (5 人の得点の
偏差積の和)}

- (5) ① 生徒 25 人の関数と図形の得点の標準偏差がともに $\sqrt{\frac{74}{25}}$ であるから、

正しい。

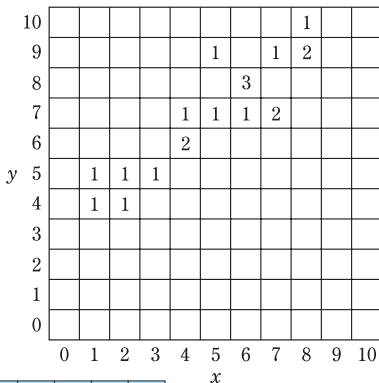
- ② 生徒 20 人のときの相関係数は 0.80, 生徒 25 人のときの相関係数は 0.76 であり, 正の相関は少し弱まるから, 正しい。

- ③ 図形の得点が一様に 1 点上がっても, 図形の標準偏差 s_y の値, 共分散 s_{xy} の値が変わらないので, 関数と図形の得点の相関係数 R の値は変わらず, 正しくない。

練習
153

右の図は, 生徒 20 人に行った漢字の読みと書きのテスト(ともに 10 点満点)の結果をまとめたものである。読みの得点 x を横軸に, 書きの得点 y を縦軸にとっている。図の中の数値は x, y の値の組に対応する人数を表している。

- (1) 各生徒の得点について, $x+y$ の最大値と, $|x-y|$ の最大値を求めよ。
(2) 図をもとに, 次の表を完成させよ。また, 各テストの得点の平均値を求めよ。



得点(点)	0	1	2	3	4	5	6	7	8	9	10
読み(人)											
書き(人)											

- (3) (2)の表を使って各テストの標準偏差を求めると, 読みは $\sqrt{5}$ 点, 書きは $\sqrt{3}$ 点で, 読みと書きの得点の共分散は 3.45 であった。読みと書きの得点の相関係数 r の値を四捨五入して小数第 2 位まで求めよ。ただし, $\sqrt{15}=3.873$ とする。

- (4) 右の表は, 別の 5 人の生徒 A, B, C, D, E に同じ問題のテストを行った結果である。5 人の読みと書きの得点の平均値は, それぞれ 20 人の得点の平均

5 人の生徒	A	B	C	D	E
読みの得点	4	2	2	8	9
書きの得点	8	4	6	7	10

値と同じであった。20 人にこの 5 人を加えた合計 25 人の生徒に関する読みと書きの得点の相関係数 R の値を四捨五入して小数第 2 位まで求めよ。ただし, $\sqrt{5}=2.236$ とする。

- (5) これらのテストの結果について, 次の①~③は正しいといえるか。

- ① 生徒 25 人の得点について, 読みと書きの平均値からの散らばり具合は同じである。
② 生徒 20 人の読みと書きの得点の正の相関はやや強いが, A~E の 5 人が加わると正の相関は少し弱まる。
③ 生徒 25 人の読みの得点が一様に 1 点下がれば, 25 人の読みと書きの得点の相関係数 R の値はより大きくなる。

→ p. 312 ⑬ ⑭

例題 154 仮説検定の考え方

赤玉と白玉があわせて10個入った袋がある。袋の中には、一方の色の玉が2個、もう一方の色の玉が8個というのはわかっているが、どちらの色の玉が何個かはわからない。この袋から花子さんが玉を1個ずつ、合計2個取り出したところ、2個とも赤玉であった。このことから花子さんは「袋の中には赤玉が8個、白玉が2個入っている」という仮説Aを立てた。取り出した玉は元に戻さないとし、起こる割合が5%以下であればほとんど起こりえないと判断するものとするとき、仮説検定により、花子さんの仮説Aが正しいか検証せよ。

考え方



- ・割合が5%以下…仮説Bは起こりえない→仮説Aが正しいと判断する
- ・割合が5%より大きい…仮説Bは起こりえる→仮説Aは正しいとはいえない

のように考える。ここで、割合が5%より大きいときも、仮説Aが正しいとはいえないだけで、正しくないわけではないことに注意する。

解答

「袋の中には赤玉が2個、白玉が8個入っている」と仮定し、これを仮説Bとする。

袋の中の2個の赤玉を R_1, R_2 ,

8個の白玉を W_1, W_2, \dots, W_8

と名前をつける。2個の玉の取り出し方は、下の表のように90通りあり、その中で、2個とも赤玉である確率は、

$$\frac{2}{90} = \frac{1}{45} = 0.022\cdots$$

したがって、2個とも赤玉である確率は約2.2%で、5%より小さいから、この仮説Bは棄却される。

よって、花子さんの仮説Aが正しいと判断される。

	R_1	R_2	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
R_1		○								
R_2	○									
W_1										
W_2										
W_3										
W_4										
W_5										
W_6										
W_7										
W_8										

▶ 仮説Bは、仮説Aの否定

▶ 数Aの「確率」の考えを使うと、赤玉2個、白玉8個から2個の玉の取り出し方は ${}_{10}C_2$ 通り、2個とも赤玉であるのは ${}_2C_2$ 通りより、

$$\frac{{}_2C_2}{{}_{10}C_2} = \frac{1}{45}$$

として求めてもよい。

▶ 仮説Bは起こりえない

↓

仮説Aが正しいと判断する

【注】例題 154 の仮説 A の否定の仮説 B を帰無仮説という。

帰無仮説が 5% の割合を基準にして 5% 以下で棄却し、仮説 A が正しいと採用したが、これは仮説 A が 95% の精度で正しいと信頼できるということで、仮説 A が正しいと言いつけるわけではない。

このことは、同じように否定を仮定して矛盾を導く背理法との違いである。(次ページの Story 参照)

【注】例題 154 の解は数え上げることで、2 個の取り出す玉の色の割合を考えたが、解の側注のように数学 A の確率の考え (p. 382 参照) を用いてもよい。

では、例題 154 では、「取り出した玉を元に戻さない」場合について考えたが、「取り出した玉を元に戻す」場合について、花子さんの仮説 A が正しいかを検証してみよう。ここでは確率の考えを利用する。(取り出した玉を元に戻す場合の検証)

帰無仮説を「袋の中には赤玉が 2 個、白玉が 8 個入っている」とする。2 個の玉の取り出し方は 100 通りで、2 個とも赤玉である取り出し方は $2 \times 2 = 4$ (通り) であるから、2 個とも赤玉である確率は、

$$\frac{4}{100} = 0.04 = 4 (\%)$$

よって、帰無仮説は棄却され、仮説 A が正しいと判断される。

【注】例題 154 では、

・袋の中には赤玉が 8 個、白玉が 2 個 ……①

・袋の中には赤玉が 2 個、白玉が 8 個 ……②

の場合では、感覚的に①の方があり得ると考えられるので、①を仮説 A として、その帰無仮説 (ここでは②) を棄却することを考えた。

一方、②を仮説 A として、その帰無仮説 (ここでは①) を考えた場合、取り出した玉が

$$2 \text{ 個とも赤である確率は、} \frac{{}_8C_2}{{}_{10}C_2} = \frac{28}{45} = 0.622 \dots$$

となり、棄却できず、②が正しいとはいえないことまではわかるが、②が正しくないとはいえない。

練習
154

袋の中に赤玉と白玉が合計 7 個入っている。

袋の中には一方の色が 2 個、他方の色が 5 個入っていることがわかっている。

A さんが袋の中から 1 個ずつ、2 回取り出したところ、2 回とも赤玉であった。

A さんはこのことから「袋の中には赤玉が 5 個、白玉が 2 個入っている」と思った。このことを検証したい。

(1) 帰無仮説を答えよ。

(2) 赤玉を 2 個取り出す確率が 5% 以下であれば帰無仮説を棄却するとして、玉の取り出し方が次の (i)、(ii) の場合について検証せよ。

(i) 1 回取り出すごとに取り出した玉を袋に戻さない場合

(ii) 1 回取り出すごとに取り出した玉を袋に戻す場合

→ p. 313 [15]

Story ストーリー

仮説検定とは？

走り高跳びの選手の太郎さんは、今までは 180 cm の高さを跳べる確率は、 $\frac{1}{2}$ であった。しかし、トレーニング方法を変えてみたところ、実力が上がったような気がした。そこで、顧問の先生にそのことを伝えると、
「今から 180 cm の高さを 8 回跳んで、7 回以上成功したら、実力が上がったと認める」と言われた。実際に、太郎さんは、8 回中 7 回成功し、顧問の先生に実力向上を認めてもらった。

このとき、太郎さんの実力が上がったといってもよいだろうか。6 回成功では実力向上とはいえないだろうか。または、8 回すべて成功しないと認められないということはないだろうか。

次のような場合について確率を用いて考え、太郎さんの実力向上について考察してみよう。（ p . 400 の反復試行の確率を利用している。）

まず、もともとの実力（180 cm の高さを跳べる確率 $\frac{1}{2}$ ）の場合に、8 回中 7 回以上成功することは難しいだろうか。7 回以上跳べる確率を求めてみよう。

【問題】

- (1) 1 回の試行で成功する（180 cm の高さを跳べる）確率が $\frac{1}{2}$ のとき、失敗する

確率も $\frac{1}{2}$ であるから、8 回跳んで 1 回も成功しない確率は、 $\left(\frac{1}{2}\right)^8 = \frac{1}{256}$

同様に考えて、8 回跳んで 8 回成功する確率も、 $\left(\frac{1}{2}\right)^8 = \frac{1}{256}$

では、8 回跳んで 7 回成功する確率はいくつだろうか。

8 回跳んで 7 回成功するということは、失敗するのは 1 回である。その 1 回の失敗が何回目に出るかを数え上げると、8 通りある。

よって、8 回跳んで 7 回成功する確率は、 $\frac{8}{256}$

1 回～7 回についても同様に考えると、成功する確率は次の表のようになる。

成功回数	0	1	2	3	4	5	6	7	8
確率	$\frac{1}{256}$	$\frac{8}{256}$	$\frac{28}{256}$	$\frac{56}{256}$	$\frac{70}{256}$	$\frac{56}{256}$	$\frac{28}{256}$	$\frac{8}{256}$	$\frac{1}{256}$

- (2) では、(1)の表より、8回跳んで7回以上成功する確率を求めると次のようになる。

$$\frac{8}{256} + \frac{1}{256} = \frac{9}{256} \div 0.035 \text{ (約 3.5\%)}$$

もし、太郎さんの実力向上が認められない、つまり、「7回以上成功したとしても実力が向上したとはいえない」と仮定すると、上のように8回跳んで7回以上成功する確率は約3.5%で、これは確率としては明らかに低い。

つまり、7回以上成功するには、1回の試行での成功確率が $\frac{1}{2}$ より高くなっていないと、なかなかできないことというのがわかる。

このことから、「実力向上が認められない」という仮定がおかしいことがわかるので、太郎さんの「実力向上が認められる」こととなる。

- (3) では、もし顧問の先生が「8回跳んで6回以上成功したら実力向上を認める」と言った場合、どうだろうか。

(1)の表より、8回跳んで6回以上成功する確率は、

$$\frac{28}{256} + \frac{8}{256} + \frac{1}{256} = \frac{37}{256} \div 0.145 \text{ (約 14.5\%)}$$

これは、7回チャレンジすれば一度は6回以上跳べる確率と考えられる。つまり、実力向上していなくても、6回以上跳べるのは無視できない確率なので、「実力向上を認める」としてしまっはいけないと考えられる。

ただし、(ここで言えるのは、「実力向上したとはいえない」だけであり、「実力向上していない」ことが言えるわけではない。)

(2)では3.5%を低い確率として、(3)では14.5%を起こりえる確率としている。このように判断するためには、本来は、『どの確率よりも低かったら「実力向上は認められない」のかを事前に定めておく』ということが必要である。その基準としてもよく用いられるのが5% (他にも10%や1%にとる場合もある) である。そのため(2)は低い確率、(3)は起こりえる確率として判断している。

このように、「あること」を示すために、その否定を仮定し、それがあらかじめ決めた基準を満たさないことを示すことで、矛盾を生じさせ、もとの「あること」を示す考え方を、「仮説検定の考え方」という。仮定したことの矛盾を示してもとのことを証明する方法には、第3章の集合と命題で学習した「背理法」があるが、背理法と仮説検定の考え方の違いは次のようになる。

【背理法と検定】

背理法

① 命題 P を証明したい② 命題 P の否定 \bar{P}
を仮定する③ \bar{P} のもとで考える

矛盾を導く

 \bar{P} は成り立たない

確率 0

④ 命題 P が示された

検定

① 仮説 H_1 を証明したい② 仮説 H_1 の否定 H_0
を仮定する③ H_0 のもとでの事象 C の
起こる確率 $P(C)$ を考える $P(C)$ が基準より
小さい

確率小

 $P(C)$ が基準より
大きい④ H_1 を採択④' H_0 を受容する

H_1 が正しいと判断される
(H_1 が証明されたわけではない.)

H_1 が正しいと判断できないだけ
で、否定されたわけではない.

では、ここでもう一度、太郎さんの走り高跳びの実力向上について考えてみよう。実際に、太郎さんは実力向上していて、1回跳んで成功する確率が0.5から0.7に上がっていたとする。そして、もし顧問の先生が「8回中8回跳べば実力向上を認める」といった場合はどうだろうか。

太郎さんが8回跳んで8回成功する確率は、 $(0.7)^8 \approx 0.058$

つまり、太郎さんは実力向上しているに関わらず、成功する確率は約5.8%となることから、この基準は厳しすぎると判断される。

今回の太郎さんの実力向上を測る基準として、顧問の先生が「8回跳んで7回以上成功する」としたことが適していることがわかるだろう。

Step Up

データの整理と分析 ▶▶ 解答編 p. 242

**

1

←
p. 282
p. 286

右の度数分布表は、ある植物の種子 100 個の重さを測定した結果である。

- (1) この度数分布表から、階級ごとの累積相対度数分布表を作成せよ。
- (2) (1)の累積相対度数分布表を用いて、累積相対度数の折れ線グラフをかけ。ただし、 x g 未満の累積相対度数 y に対して、点 (x, y) をとること。
- (3) (2)のグラフを利用して、この種子 100 個の重さの中央値を求めよ。

階級 (g)	度数 (個)
3 以上 4 未満	2
4 以上 5 未満	10
5 以上 6 未満	13
6 以上 7 未満	20
7 以上 8 未満	25
8 以上 9 未満	12
9 以上 10 未満	11
10 以上 11 未満	7
合計	100

**

2

←
p. 288
p. 290

次の表は、ある試験とその追試の結果である。試験後には、直しのレポートを提出させ、よく復習してから追試を受けるように指示した。箱ひげ図をかき、それを比較し、復習はよくなされたといえるかどうか答えよ。

得点 (点)	0	1	2	3	4	5	6	7	8	9	10	合計
試験 (人)	0	0	0	1	3	2	6	4	6	2	1	25
追試 (人)	0	0	0	0	1	3	2	4	6	7	2	25

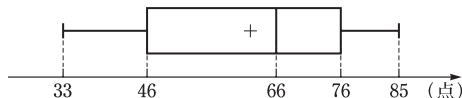
**

3

←
p. 288

次の表は、生徒 10 人のテストの点数で、下はその箱ひげ図である。平均値が 62 点であるとき、表の a , b , c , d の値をそれぞれ求めよ。ただし、 $a < b < c < d$ とする。

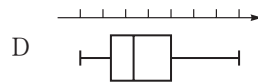
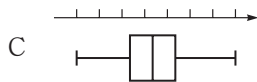
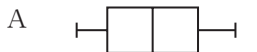
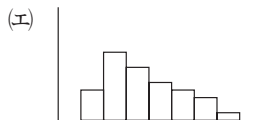
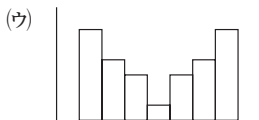
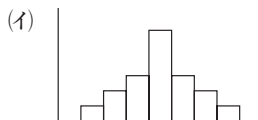
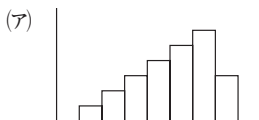
出席番号	1	2	3	4	5	6	7	8	9	10
点数 (点)	42	33	a	52	b	d	76	c	82	68



**

4

次の(ア)～(エ)のヒストグラムについて、同じデータを使って表した箱ひげ図としてふさわしいものを、A～Dから選べ。

←
p. 291

**

5

8 個のデータの値 x_1, x_2, \dots, x_8 の平均値は 3, 分散は 4 である.

←
p. 292

これに $x_9=6, x_{10}=5$ をつけ加えた 10 個のデータの値 x_1, x_2, \dots, x_{10} の平均値と分散を求めよ.

**

6

10 個のデータの値がある. そのうちの 6 個のデータの値の平均値は 3, 分散は 9 である. 残りの 4 個のデータの値の平均値は 8, 分散は 14 である. この 10 個のデータの値の平均値と分散を求めよ.

←
p. 292

7

任意の連続する 5 個の自然数の分散 s^2 を求めよ.

←
p. 292

8

変数 x の n 個のデータの値 x_1, x_2, \dots, x_n がある. x の平均値が a (定数) であるとき, このデータ x の分散 s^2 の最小値を求めよ.

←
p. 292

9

←
p. 294

変量 x の n 個のデータの値 x_1, x_2, \dots, x_n がある. x の平均値 \bar{x} が 6, 標準偏差 s_x が $2\sqrt{2}$ であるとき, 変量 $y = \frac{1}{4}x + 20$ の平均値 \bar{y} と標準偏差 s_y を求めよ.

10

←
p. 296

次の表は, 高校生の兄弟 9 組の身長を計測した結果である.

番号	1	2	3	4	5	6	7	8	9
兄の身長 (cm)	166	170	179	173	184	172	169	163	172
弟の身長 (cm)	165	170	a	174	176	171	166	166	167

弟の身長の平均値は, 兄の身長の平均値より 2 cm 小さいという.

- (1) 変量 x と変量 u の間に $x = mu + n$ (m, n は定数) という関係があるとき, x の平均値 \bar{x} と u の平均値 \bar{u} の間には $\bar{x} = m\bar{u} + n$ という関係が成り立つ. 兄の身長 x cm を $x = u + 170$ と考えて, その平均値 \bar{x} を求めよ.
- (2) a の値を求めよ.
- (3) 兄の身長 x cm と弟の身長 y cm の散布図をかけ.
- (4) 兄の身長 x cm と弟の身長 y cm の相関係数 r を求めよ. また, 兄の身長と弟の身長の間には, 相関関係があるといえるか.

第5章

11

←
p. 296

2 つの変量 x, y をもつ n 個のデータ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ がある. x と y の間に $y = 7x + 8$ の関係があるとき, x と y の相関係数 r を求めよ.

**

12

←
p. 296

右の表は, 芸術大学の学生 10 人がそれぞれ作品 A (絵画) と作品 B (彫刻) を制作し, それらに対して 0, 1, 2 の 3 段階で評価を行った際の得点を 2 次元の度数分布表にまとめたものである. A の得点 x と B の得点 y の相関係数 r を小数第 2 位まで求めよ. ただし, $\sqrt{41} = 6.403$ とする.

B \ A	0	1	2	合計
0	0	1	0	1
1	1	3	1	5
2	0	2	2	4
合計	1	6	3	10

**

13

←
p. 301

99 個の観測値からなるデータがある。次の四分位数について書かれた文章のうち、どのようなデータでも成り立つものをすべて選べ。

- ① 平均値は第 1 四分位数と第 3 四分位数の間にある。
- ② 四分位範囲は標準偏差より大きい。
- ③ 中央値より小さい観測値の個数は 49 個である。
- ④ 最大値に等しい観測値を 1 個削除しても第 1 四分位数は変わらない。
- ⑤ 第 1 四分位数より小さい観測値と、第 3 四分位数より大きい観測値とをすべて削除すると、残りの観測値の個数は 51 個である。
- ⑥ 第 1 四分位数より小さい観測値と、第 3 四分位数より大きい観測値とをすべて削除すると、残りの観測値からなるデータの範囲はもとのデータの四分位範囲に等しい。

(2020 センター試験・改)

14

←
p. 301

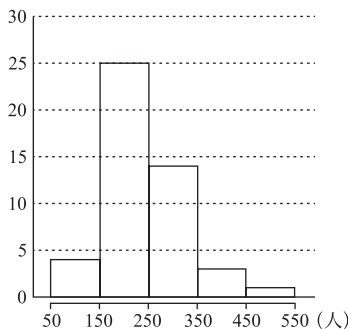
右の図は、2008 年における 47 都道府県の人口 1 万人あたりの旅券取得者数のヒストグラムである。なお、ヒストグラムの各階級の区間は、左側の数値を含み、右側の数値を含まない。

このヒストグラムに対して、各階級に含まれるデータの値がすべてその階級値に等しいと仮定するとき、次の問いに答えよ。

- (1) この変量の平均値 \bar{x} を小数第 1 位を四捨五入して求めよ。
- (2) この変量の分散 s^2 に値に最も近いものを次の①～⑧から 1 つ選べ。

- ① 3900 ② 4900 ③ 5900 ④ 6900
- ⑤ 7900 ⑥ 8900 ⑦ 9900 ⑧ 10900

(都道府県数)



2008 年における旅券取得者数のヒストグラム

(出典：外務省の Web ページにより作成)

(2021 大学入学共通テスト・改)

15

次の問いに答えよ。

p. 304

- (1) ある硬貨Aがある。恵さんが、この硬貨Aを投げて表が出るか裏が出るかという実験を行い、次のような結果が得られた。

恵さん：硬貨Aを10回投げて、表が7回出た。
この結果から、「硬貨Aは表が出やすく加工されている」と恵さんは考えた。

起こる割合が5%以下であればほとんど起こりえないと判断するものとし、恵さんの考えが正しいと判断できるか答えよ。なお、硬貨を10回投げたときの表裏の出方は1024通りで、表が7回以上出る出方は176通りである。

- (2) 今度は、光さんが(1)の硬貨Aを投げて表裏の出方の実験を行ったところ、100回中70回表が出た。100回投げて70回以上の表が出る確率を計算すると、約0.004%になることから、光さんは、「硬貨Aは表が出やすく加工されている」と考えた。

恵さんと光さんの考えを利用して、次の文の内容が正しいと判断できるかどうかを考えた。(ア)、(イ)の○×の組合せが正しいものを、下の選択肢①～④より1つ選べ。

- (ア) 1人の当選者を選ぶ選挙に、S候補とT候補の2名が立候補した。第1投票所での出口調査で無作為に選んだ10人のうち7人がS候補に投票したので、S候補が当選すると判断した。
- (イ) 新発売の商品のパッケージデザインをA案にするかB案にするかの消費者アンケートを100人選んで行い、投票数の多いほうの案を消費者全体に好まれそうな案だと判断した。

	①	②	③	④
(ア)	○	○	×	×
(イ)	○	×	○	×

○：正しいと判断できる，×：正しいと判断できない

章末問題

▶▶ 解答編 p. 255

1 ←
p. 304 新商品のパッケージデザインとしてAとBの2案を作成し、街頭アンケートで消費者30人に好みを選んでもらったところ、B案を選んだ人が23人いた。

一般に、B案の方が消費者に好まれる案だと判断してもよいだろうか。もしA案とB案を何も見ずに選ぶ場合、それぞれが選ばれる確率は0.5とし、起こる割合が5%以下であればほとんど起こりえないと判断するものとする。

また、下の表は、表と裏が同じ割合で出るコイン30枚を同時に投げたときの、表が出た枚数を記録することを1000回行った結果である。必要に応じて用いてもよい。

表の枚数	回	表の枚数	回
0~15	526	23	5
16	132	24	4
17	131	25	3
18	95	26	3
19	46	27	1
20	32	28	0
21	14	29	0
22	8	30	0

思考力問題

2 総務省が実施している国勢調査では都道府県ごとの総人口が調べられており、その内訳として日本人人口と外国人人口が公表されている。また、外務省では旅券(パスポート)を取得した人数を都道府県ごとに公表している。加えて、文部科学省では都道府県ごとの小学校に在籍する児童数を公表している。

そこで、47都道府県の、人口1万人あたりの外国人人口(以下、外国人数)、人口1万人あたりの小学校児童数(以下、小学生数)、また、日本人1万人あたりの旅券を取得した人数(以下、旅券取得者数)を、それぞれ計算した。

次の図1は、2010年における47都道府県の、旅券取得者数(横軸)と小学生数(縦軸)の関係を黒丸で、また、旅券取得者数(横軸)と外国人数(縦軸)の関係を白丸で表した散布図である。

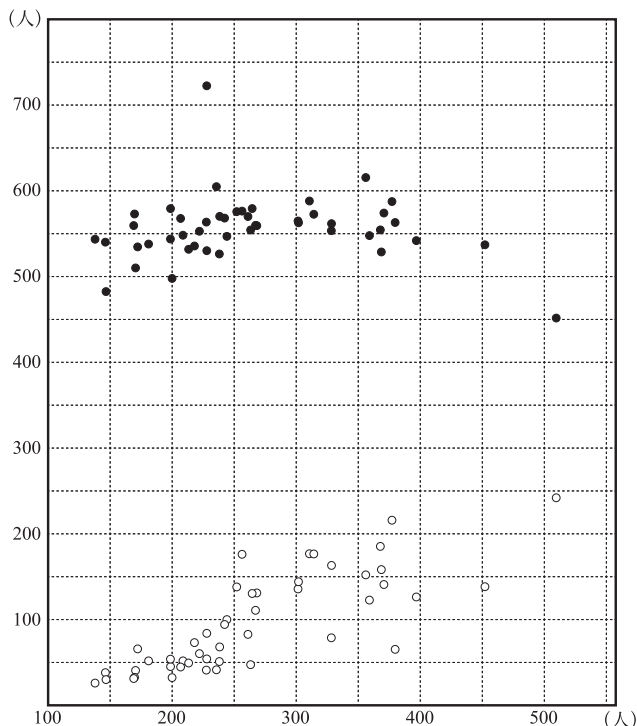


図1 2010年における、旅券取得者数と小学生数の散布図(黒丸)、
旅券取得者数と外国人数の散布図(白丸)
(出典：外務省、文部科学省および総務省のWebページにより作成)

次の(I)、(Ⅱ)、(Ⅲ)は図1の散布図に関する記述である。

(I)～(Ⅲ)の文が正しいかどうか答えよ。

- (I) 小学生数の四分位範囲は、外国人数の四分位範囲より大きい。
- (Ⅱ) 旅券取得者数の範囲は、外国人数の範囲より大きい。
- (Ⅲ) 旅券取得者数と小学生数の相関係数は、旅券取得者数と外国人数の相関係数より大きい。

(2021 大学入学共通テスト・改)