1節　データの分析

1 データと度数分布表

ねらい　データは，整理することによって，全体の傾向や特徴を見やすくすることができます。データを整理する方法を学びます。

下の資料は，あるクラスを，電車で通学する生徒 20 人の A 班と，徒歩で通学する生徒 15 人の B 班に分けて，先月の読書時間の合計を調べたものである。

A 班　　　　　　　　　　　　　　（単位　時間）

3	10	7	14	5	9	15	0	9	18
0	8	11	10	15	19	6	23	13	5

B 班　　　　　　　　　　　　　　（単位　時間）

6	20	0	14	16	23	1	4	5	0
18	13	21	0	9					

このような資料を**データ**という。データは個々の値を並べただけでは，全体の傾向や特徴はわからない。特徴を調べるときは，目的に合わせて整理することが大切である。

度数分布表とヒストグラム

左下の表は，A 班のデータを，時間を 4 時間ずつの区間に分け，その区間に入る人数を調べてまとめたものである。

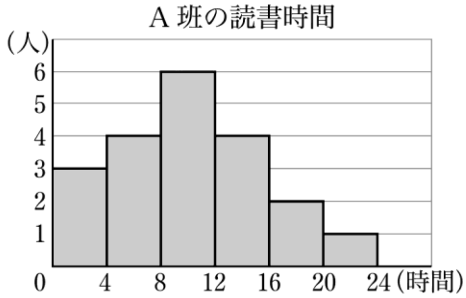
各区間を**階級**，各階級に入っているデータの値の個数を**度数**，階級の中央の値を**階級値**という。このように，各階級に度数を対応させた表を**度数分布表**という。

右下の図は，分布のようすを見やすくするために，A 班の読書時間の度数分布表から，階級の幅を底辺，度数を高さとする長方形をすきまなく並べたグラフである。このようなグラフを**ヒストグラム**という。

A 班の読書時間

時間の階級 （時間）	階級値 （時間）	度数 （人）
0 以上～ 4 未満	2	3
4 ～ 8	6	4
8 ～ 12	10	6
12 ～ 16	14	4
16 ～ 20	18	2
20 ～ 24	22	1
計		20

A 班の読書時間



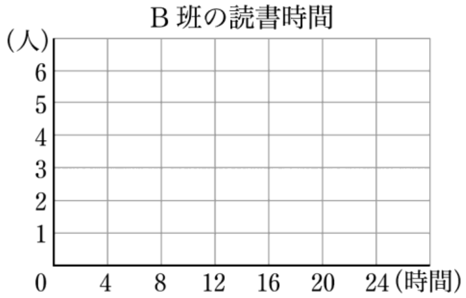
度数分布表やヒストグラムに表すことで，A 班では，8 時間以上 12 時間未満の階級の度数が最も大きいことがわかる。

問 1 128 ページの B 班の読書時間を度数分布表にまとめ，ヒストグラムに表しなさい。また，度数の最も大きい階級を答えなさい。

B 班の読書時間

時間の階級（時間）	階級値（時間）	度数（人）
0 以上～ 4 未満	2	
4 ～ 8	6	
8 ～ 12	10	
12 ～ 16	14	
16 ～ 20	18	
20 ～ 24	22	
計		

B 班の読書時間



相対度数

A 班と B 班では，度数の合計が異なるので，各階級の度数をそのまま比べても違いがはっきりしない。

度数の合計が異なるときは，各階級の度数が，度数の合計に対してどのような割合であるか調べてみるとよい。

各階級の度数を度数の合計でわった値を**相対度数**といい，各階級に相対度数を対応させた表を**相対度数分布表**という。

A 班の読書時間の相対度数分布表は，右のようになる。

A 班，B 班の読書時間

時間の階級（時間）	A 班		B 班	
	度数（人）	相対度数	度数（人）	相対度数
0 以上～ 4 未満	3	0.15		
4 ～ 8	4	0.20		
8 ～ 12	6	0.30		
12 ～ 16	4	0.20		
16 ～ 20	2	0.10		
20 ～ 24	1	0.05		
計	20	1.00		

問 2 上の B 班の読書時間の相対度数分布表を完成しなさい。また，16 時間以上 20 時間未満の階級では，A 班と B 班で，どちらの相対度数が大きいか答えなさい。

2 代表値

ねらい　データ全体の特徴を表す数値の求め方や選び方について学びます。

データ全体の特徴を表す数値を**代表値**という。代表値には，平均値，中央値，最頻値などがある。

平均値と中央値

よく知られた代表値に平均値がある。**平均値**は，データの値の合計をデータの値の個数でわった値である。

◀ 平均値 $= \dfrac{\text{データの値の合計}}{\text{データの値の個数}}$

● A 班の読書時間の平均値を求めてみよう。

例 1　$\dfrac{3+10+7+\cdots+13+5}{20} = \dfrac{200}{20} = 10$（時間）

問 3　128 ページの B 班の読書時間の平均値を求めなさい。

データの値を小さい順に並べたとき，中央の値を**中央値**という。ただし，データの値の個数が偶数のときは，中央にある 2 つの値の平均値を中央値とする。

◀中央値はメジアンともいう。

◀奇数のとき

○○●○○

中央値

偶数のとき

○○●●○○

この 2 つの値の平均値が中央値

● A 班の読書時間の中央値を求めてみよう。

例 2　データの値を小さい順に並べると，次のようになる。

$$\underbrace{0\ 0\ 3\ 5\ 5\ 6\ 7\ 8\ 9\ 9}_{10\text{人}}\ \underbrace{10\ 10\ 11\ 13\ 14\ 15\ 15\ 18\ 19\ 23}_{10\text{人}}$$

中央値は，10 番目と 11 番目の平均値である。

よって　$\dfrac{9+10}{2} = 9.5$（時間）

問 4　128 ページの B 班の読書時間を，値の小さい順に並べると次のようになる。B 班の読書時間の中央値を求めなさい。

➔p.140 復習問題①

0　0　0　1　4　5　6　9　13　14　16　18　20　21　23

データのなかに，ほかの値とかけ離れた値がある場合がある。このようなときは，平均値よりも中央値の方がその値の影響を受けにくいので，代表値としては，平均値より中央値が適当である。

最頻値

　度数分布表で，度数が最も大きい階級の階級値を**最頻値**という。最頻値はヒストグラムにおいて，最も高い長方形に対応する階級値である。

◀最頻値はモードともいう。また，データにおいて個数が最も多い値を最頻値ということもある。

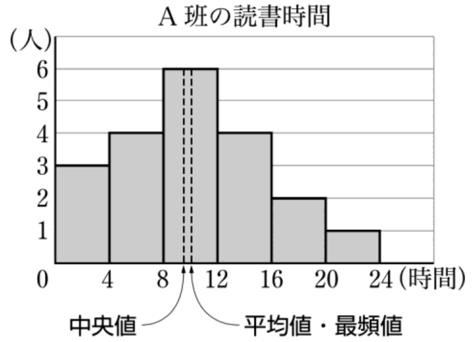
● A 班の読書時間の最頻値を求めてみよう。

例 3　128 ページの A 班の読書時間の度数分布表で，度数が最も大きい階級は 8 時間以上 12 時間未満である。したがって，最頻値はその階級値の 10 時間である。

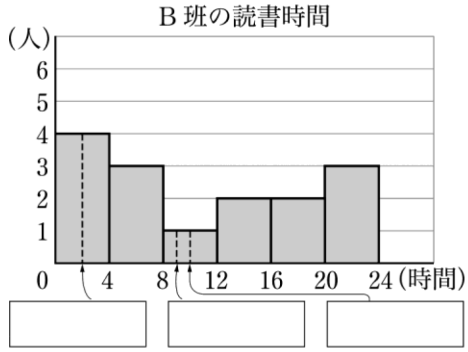
問 5　129 ページの問 1 でつくった度数分布表から, B 班の読書時間の最頻値を求めなさい。

● A 班と B 班について，読書時間の 3 つの代表値を比べてみよう。

例 4　A 班の読書時間の 3 つの代表値をヒストグラムに表すと，次の図のようになる。



問 6　例 4 にならって，B 班の読書時間の 3 つの代表値を，次のヒストグラムの［　　　］に入れなさい。



3 四分位数と箱ひげ図

ねらい　データの散らばりぐあいも，データ全体のもつ特徴の 1 つです。中央値をもとにして，データの散らばりぐあいを，数値や図で表すことを学びます。

四分位数と四分位範囲

データは，散らばりぐあいを考えることが大切な場合もある。データの散らばりぐあいは，代表値ではとらえられない。

データの値で

$$(\text{最大値})-(\text{最小値})$$

を，そのデータの分布の**範囲**という。範囲を調べて，データの散らばりぐあいを比べてみよう。

◀データの値には，かならず最大値と最小値がある。

下の表は，バスケットボール部の A さんと B さんの，最近 10 試合で成功したシュートの本数である。

A さん　（単位　本）

5	3	6	5	6	7	8	7	8	10

B さん　（単位　本）

4	5	13	5	9	6	6	7	6	5

2 人の成功したシュートの本数の範囲は，それぞれ

A さん　$10-3=7$（本）　　B さん　$13-4=9$（本）

となる。範囲の値を比べると，散らばりぐあいは A さんより B さんの方が大きいといえる。

範囲は，データのなかにほかとかけ離れた値がある場合，その値の影響を受けやすい。その影響を少なくしたものに，**四分位範囲**がある。四分位範囲は次のようにして求める。

① データの値を小さい順に並べ，中央値を境にして 2 つに分ける。

② 中央値を，**第 2 四分位数**という。

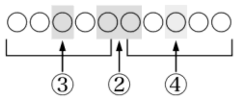
③ 最小値を含む方のデータの中央値を**第 1 四分位数**という。

④ 最大値を含む方のデータの中央値を**第 3 四分位数**という。

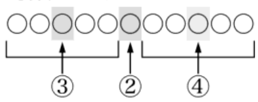
⑤ $(\text{第 3 四分位数})-(\text{第 1 四分位数})$ が，四分位範囲である。

また，四分位範囲を 2 でわった値を**四分位偏差**という。

◀偶数のとき



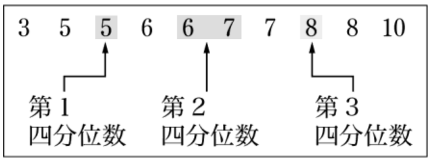
奇数のとき



◀第 3 四分位数と第 1 四分位数の間に，データの個数のほぼ半分が含まれる。

●Aさんの成功したシュートの本数の四分位範囲を求めてみよう。

例 5　Aさんの成功したシュートの本数を，小さい順に並べかえると右のようになるから



第2四分位数　$\frac{6+7}{2}=6.5$（本）
第1四分位数　5本
第3四分位数　8本
四分位範囲　$8-5=3$（本）
四分位偏差　$3 \div 2 = 1.5$（本）

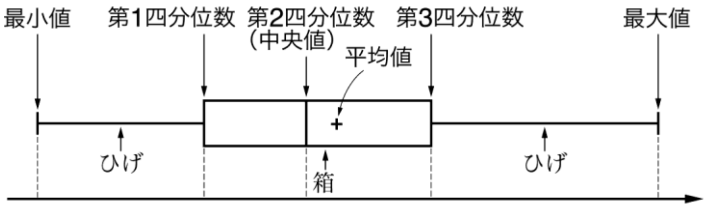
◀四分位偏差 $= \dfrac{\text{四分位範囲}}{2}$

問 7　Bさんの成功したシュートの本数の，四分位範囲，四分位偏差を求めなさい。また，Aさんと Bさんの四分位範囲を比べて，どちらの散らばりぐあいが大きいか答えなさい。

箱ひげ図

四分位数を用いて，データの散らばりぐあいを見やすく表すには，下のように長方形に線を添えた**箱ひげ図**を用いる。

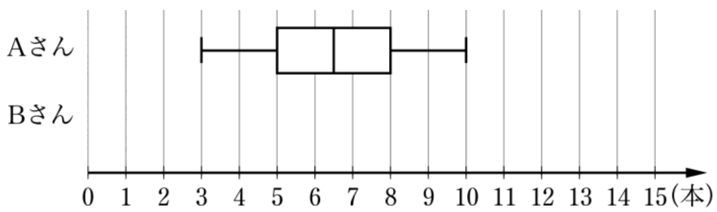
◀第1四分位数，第2四分位数，第3四分位数をまとめて，四分位数という。



◀平均値を表す＋をかかないこともある。

●Aさんの成功したシュートの本数の箱ひげ図をつくってみよう。

例 6　例5の結果から，箱ひげ図は次のようになる。



問 8　Bさんの成功したシュートの本数の箱ひげ図を，上の図に表しなさい。

→p.140 復習問題2

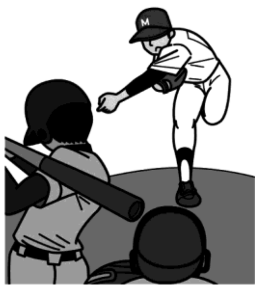
4　分散と標準偏差

ねらい　データの散らばりぐあいを，平均値をもとにして，1つの数値で表すことを学びます。

データの散らばりぐあいを, データの個々の値と平均値との差を用いて考えてみよう。

(データの個々の値)−(平均値)

を**偏差**という。



次の表は，ある野球チームのA投手とB投手が，最近5試合で奪った三振の個数である。

（単位　個）

	1試合目	2試合目	3試合目	4試合目	5試合目
A投手	4	8	7	5	6
B投手	10	2	9	6	3

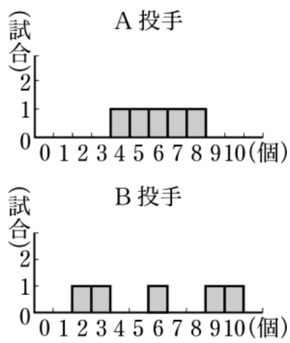
平均値は，それぞれ

A投手　$\frac{4+8+7+5+6}{5}=\frac{30}{5}=6$（個）

B投手　$\frac{10+2+9+6+3}{5}=\frac{30}{5}=6$（個）

となり，等しくなる。しかし，三振の個数をヒストグラムに表すと，散らばりぐあいは異なる。

◀三振の個数のヒストグラム



次に，A投手，B投手の三振の個数の偏差とその合計を求めると下の表のようになる。

	1試合目	2試合目	3試合目	4試合目	5試合目	計
A投手の偏差	−2	2	1	−1	0	0
B投手の偏差	4	−4	3	0	−3	0

この表からわかるように，偏差の合計はどちらも0になるため，偏差の平均値でデータ全体の散らばりぐあいを表すことはできない。

そこで，偏差を2乗した値を考える。偏差の2乗は0以上の値になるので，合計が0に近いほど平均値からの散らばりぐあいが小さいといえる。

偏差の 2 乗の平均値を**分散**といい，$\boldsymbol{s^2}$ で表す。また，分散の正の平方根を**標準偏差**といい，$\boldsymbol{s}$ で表す。

分散と標準偏差	
偏差	(データの個々の値)−(平均値)
分散	$s^2=(\text{偏差})^2$ の平均値 $=\dfrac{(\text{偏差})^2 \text{ の合計}}{\text{データの値の個数}}$
標準偏差	$s=\sqrt{s^2}=(\text{分散の正の平方根})$

分散や標準偏差は，データ全体の散らばりぐあいを表す数値である。

分散や標準偏差の値は，0 に近いほどデータの個々の値が平均値の近くに分布していることを意味し，大きいほどデータの個々の値に平均値から離れたものが多くあることを意味している。

● A 投手の三振の個数の分散，標準偏差を求めてみよう。

例 7　134 ページの A 投手が奪った三振の個数から偏差の 2 乗の合計を計算すると，右の表のようになる。
したがって，分散 s^2 は

$$s^2=\frac{10}{5}=2$$

標準偏差 s は

$$s=\sqrt{2}=1.4142\cdots \fallingdotseq 1.41 \text{（個）}$$

A 投手	三振の個数	偏差	$(\text{偏差})^2$
1 試合目	4	−2	4
2 試合目	8	2	4
3 試合目	7	1	1
4 試合目	5	−1	1
5 試合目	6	0	0
計	30	0	10

問 9　右の表を完成して，B 投手が奪った三振の個数の分散を求めなさい。また，標準偏差を，四捨五入して小数第 2 位まで求めなさい。

B 投手	三振の個数	偏差	$(\text{偏差})^2$
1 試合目	10		
2 試合目	2		
3 試合目	9		
4 試合目	6		
5 試合目	3		
計	30		

問 10　A 投手と B 投手が奪った三振の個数の標準偏差を比べて，どちらの散らばりぐあいが大きいか答えなさい。

5 相関関係

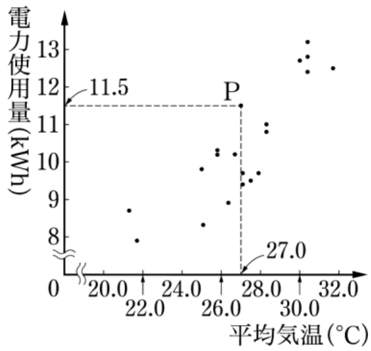
ねらい 2つの数量の関係を図を使って調べてみます。また，その関係のどあいについて考えます。

散布図

たとえば，平均気温と電力使用量のように，互いに関係があると考えられる数量がある。このような2つの数量を組にしたデータについて考えてみよう。

2つの数量の関係は，それぞれの値の組をx座標，y座標とする点として，平面上に表すことができる。このような図を**散布図**という。

右の図は，ある町の7月中の連続した20日間における日ごとの平均気温と1世帯あたりの電力使用量を散布図に表したものである。図中の点Pは，平均気温が27.0℃で，電力使用量が11.5kWhであることを表している。



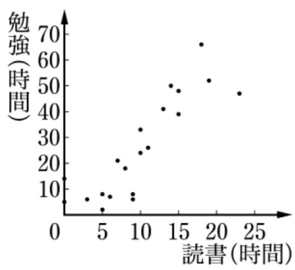
相関関係

下の表は，20人の生徒について，先月の読書時間と，勉強時間，テレビ視聴時間，1日のメール発信回数の平均を調べたデータである。

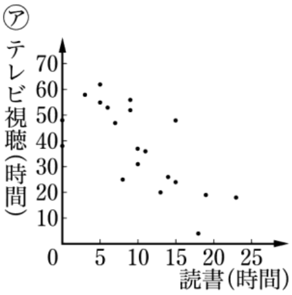
読書（時間）	3	10	7	14	5	9	15	0	9	18	0	8	11	10	15	19	6	23	13	5
勉強（時間）	6	33	21	50	2	8	39	5	6	66	14	18	26	24	48	52	7	47	41	8
テレビ視聴（時間）	58	37	47	26	55	52	48	38	56	4	48	25	36	31	24	19	53	18	20	62
メール発信（回）	5	2	9	7	8	6	8	3	2	2	2	6	12	3	9	6	11	1	4	1

読書時間と勉強時間を散布図に表すと，右の図のようになる。

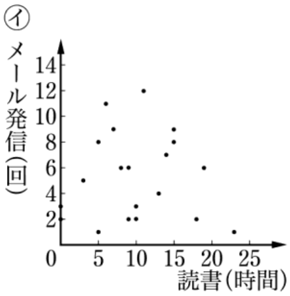
この散布図から，読書時間と勉強時間は，一方が増加すれば他方も増加する傾向があることがわかる。このとき，2つの数量の間には**正の相関関係**があるという。



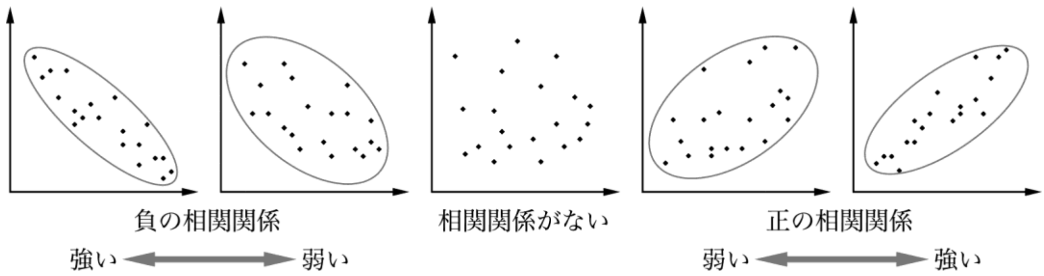
右の㋐の図は，読書時間とテレビ視聴時間の散布図である。この散布図から，読書時間とテレビ視聴時間は，一方が増加すれば他方は減少する傾向があることがわかる。このとき，2 つの数量の間には**負の相関関係**があるという。



また，右の㋑の図は，読書時間とメール発信回数の散布図である。読書時間とメール発信回数には，正の相関関係も負の相関関係もみられない。このとき，2 つの数量の間には相関関係がないという。

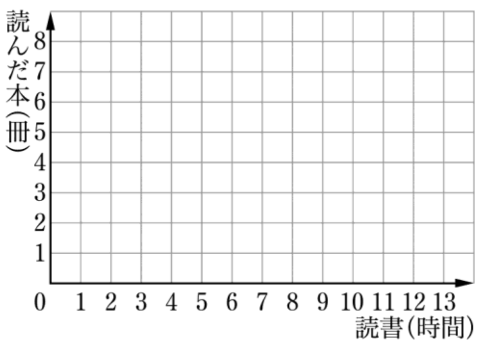


2 つの数量の間に相関関係があるとき，散布図の点が直線状に近づくほど相関関係が強いといい，直線状ではなく広く散らばるほど相関関係が弱いという。



問 11　次の表は，5 人の生徒の，先月の読書時間と読んだ本の冊数を示したものである。散布図をつくり，読書時間と読んだ本の冊数には，どのような相関関係があるか答えなさい。

生徒	読書（時間）	読んだ本（冊）
a	5	2
b	8	6
c	9	4
d	12	8
e	11	5



6　相関係数

ねらい　相関関係は，散布図でとらえることができますが，数値で表すことができれば，さらに便利です。相関関係を数値で表すことを学びます。

相関係数の意味

相関関係を調べたい2つの数量を x，y とする。x の偏差と y の偏差の積の平均値を**共分散**という。また，共分散を x の標準偏差と y の標準偏差の積でわった値を**相関係数**という。相関係数は記号 $\boldsymbol{r}$ で表す。

◀(偏差)＝(データの個々の値)－(平均値)

共分散と相関係数

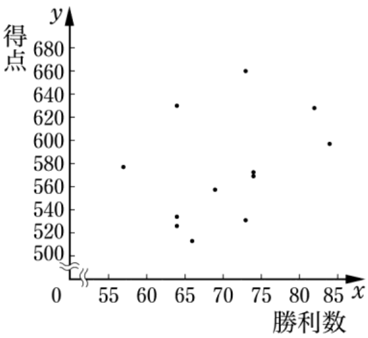
共分散　　x と y の偏差の積の平均値

相関係数　　$r=\dfrac{\text{共分散}}{(x\text{の標準偏差})\times(y\text{の標準偏差})}$

相関関係を数値で表すには，共分散や相関係数が用いられる。とくに，相関係数 r の値については，不等式

$$-1 \leqq r \leqq 1$$

が成り立ち，正の相関関係が強いほど1に近づき，負の相関関係が強いほど -1 に近づく。



プロ野球12チームについて，1年間の勝利数 x と得点 y を調べたところ，共分散は123.9，x の標準偏差は7.5，y の標準偏差は44.5であった。
これらを用いて相関係数 r は次のように求められる。

$$r=\frac{123.9}{7.5\times44.5}=0.3712\cdots \fallingdotseq 0.37$$

問12　上のプロ野球12チームで，勝利数と失点数の相関係数は -0.73，勝利数と得失点差の相関係数は0.90である。
次の①，②から，正しいものを選びなさい。

(1)　勝利数と失点数には

①正の相関関係がある　　②負の相関関係がある

(2)　勝利数と得点より，勝利数と得失点差の方が

①正の相関関係が強い　　②正の相関関係が弱い

➔p.140 復習問題3

相関係数

●読書時間と読んだ本の冊数の相関係数を求めてみよう。

例 8 137 ページの問 11 について，読書時間を x とし，読んだ本の冊数を y とする。

$(x\text{の平均値}) = \frac{5+8+9+12+11}{5} = 9$（時間）

$(y\text{の平均値}) = \frac{2+6+4+8+5}{5} = 5$（冊）

下の表のようにして計算すると

生徒	x	y	xの偏差	yの偏差	$(x\text{の偏差})^2$	$(y\text{の偏差})^2$	偏差の積
a	5	2	-4	-3	16	9	12
b	8	6	-1	1	1	1	-1
c	9	4	0	-1	0	1	0
d	12	8	3	3	9	9	9
e	11	5	2	0	4	0	0
計	45	25	0	0	30	20	20

x，y の共分散は　$\frac{20}{5} = 4$

x の標準偏差は　$\sqrt{\frac{30}{5}} = \sqrt{6}$

y の標準偏差は　$\sqrt{\frac{20}{5}} = \sqrt{4} = 2$

相関係数 r は

$$r = \frac{4}{\sqrt{6} \times 2} = \frac{\sqrt{6}}{3} = 0.816\cdots \fallingdotseq 0.82$$

◀電卓などを用いて，求めるとよい。

問 13 次の表は，4 人の生徒の数学と英語の小テストの得点を示したものである。表を完成することにより，数学の得点 x と英語の得点 y の相関係数を求めなさい。

生徒	x	y	xの偏差	yの偏差	$(x\text{の偏差})^2$	$(y\text{の偏差})^2$	偏差の積
a	6	5					
b	7	5					
c	7	8					
d	8	10					
計							

復習問題

□ 1　次の資料は，1 年生の図書委員 8 人が 1 年間に学校の図書室から借りた本の冊数を調べたものである。

代表値
p.130 例1
p.130 例2

7	12	8	40	9	4	8	8

(1)　借りた本の冊数の平均値，中央値を求めなさい。

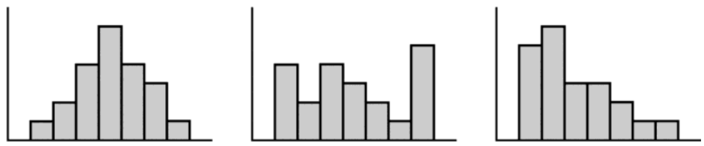
(2)　40 冊借りた生徒を除いて，残り 7 人が借りた本の冊数の平均値，中央値を求めなさい。

□ 2　次のヒストグラムについて，対応する箱ひげ図を選びなさい。

箱ひげ図
p.133 例6

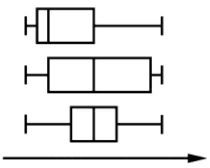
ヒストグラム

(1)　(2)　(3)



箱ひげ図

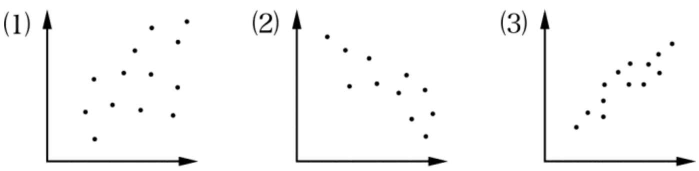
①
②
③



□ 3　次の散布図について，対応する相関係数を選びなさい。

相関関係，相関係数
p.137 問11
p.138 問12

散布図



相関係数

①　0.9　　②　0.5　　③　−0.8