データの分析

テストの点数や身長・体重など，ある集団を構成する人や物の特性を数量的に表わすものを**変量**といい，調査や実験などで得られた数値を**データ**という。
ここでは，このデータを分析するために必要な手法を学んでいく。

例えば，調査をした結果，次のようなデータが得られたとします。

① **定期考査の点数が悪い人の 98%が日常的に<u>この食べ物</u>を摂取していた。**
② **ケガをよくする人の 95%が日常的に<u>この食べ物</u>を摂取していた。**
③ **忘れ物をよくする人の 90%が日常的に<u>この食べ物</u>を摂取していた。**

さて，皆さんはこのデータを見て，「この食べ物」を禁止すべきだと思いますか？
ここで，「禁止すべきだ！」と思った人は要注意。データにだまされやすい人ですよ。
よく考えてみてくださいね。

§1　代表値

データ全体の特徴を，1 つの数値を目安として表すことがある。このような数値を**代表値**という。
以下におもな 3 つをあげる。

○平均値

変量 x がとる n 個の値 x_1，x_2，…，x_n からなる 1 組のデータにおいて，これらの値の総和を n で割ったものを平均値といい，$\bar{x}$ で表す。

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

例 1　次の数値は，高校 1 年生 10 名の身長である。平均値を求めなさい。
154.7　158.4　155.2　156.4　162.3　157.3　157.0　159.7　160.4　158.4　（単位は cm）

まずは定義通り計算していきます。

$$\frac{154.7+158.4+155.2+156.4+162.3+157.3+157.0+159.7+160.4+158.4}{10}=\frac{1579.8}{10}$$
$$=157.98\ (\mathrm{cm})$$

次に少し工夫をしてみます。ここでは，157 との差の平均値を考えていきます。

10 個のデータから 157 を引くと，

-2.3　1.4　-1.8　-0.6　5.3　0.3　0　2.7　3.4　1.4

この 10 個の値の平均値は，

$$\frac{-2.3+1.4-1.8-0.6+5.3+0.3+0+2.7+3.4+1.4}{10}=\frac{9.8}{10}=0.98$$

これが，157 との差の平均値を表しているので，求める平均値は，

$157+0.98=157.98$ (cm)

このように計算すると，各値が小さくなり，かつ正負で打ち消しあうので計算しやすくなる。
ここで用いた 157 を**仮平均**と言う。なお，仮平均が本当の平均と一致していた場合，差の平均値はちょうど 0 になる。

○中央値（メジアン）

データを大きい順に並べたとき，その中央の値のことを**中央値**または**メジアン**と言う。
データの個数が奇数のときは，そのまま中央の値をとればよいが，偶数のときは中央に 2 つの値が並ぶので，このときはその 2 つの値の平均を考える。

データA
168.5
160.4
159.0
155.3
149.1

→ **中央値** 159.0(cm)

データB
168.5
160.4
159.0
156.8
155.3
149.1

→ **中央値** $\frac{159.0+156.8}{2}=157.9$ (cm)

○最頻値（モード）

データにおいて最も個数の多い値を**最頻値**または**モード**という。

例 2　例 1 のデータにおいて，中央値と最頻値を求めなさい。

例 1 のデータを小さい順に並べると，

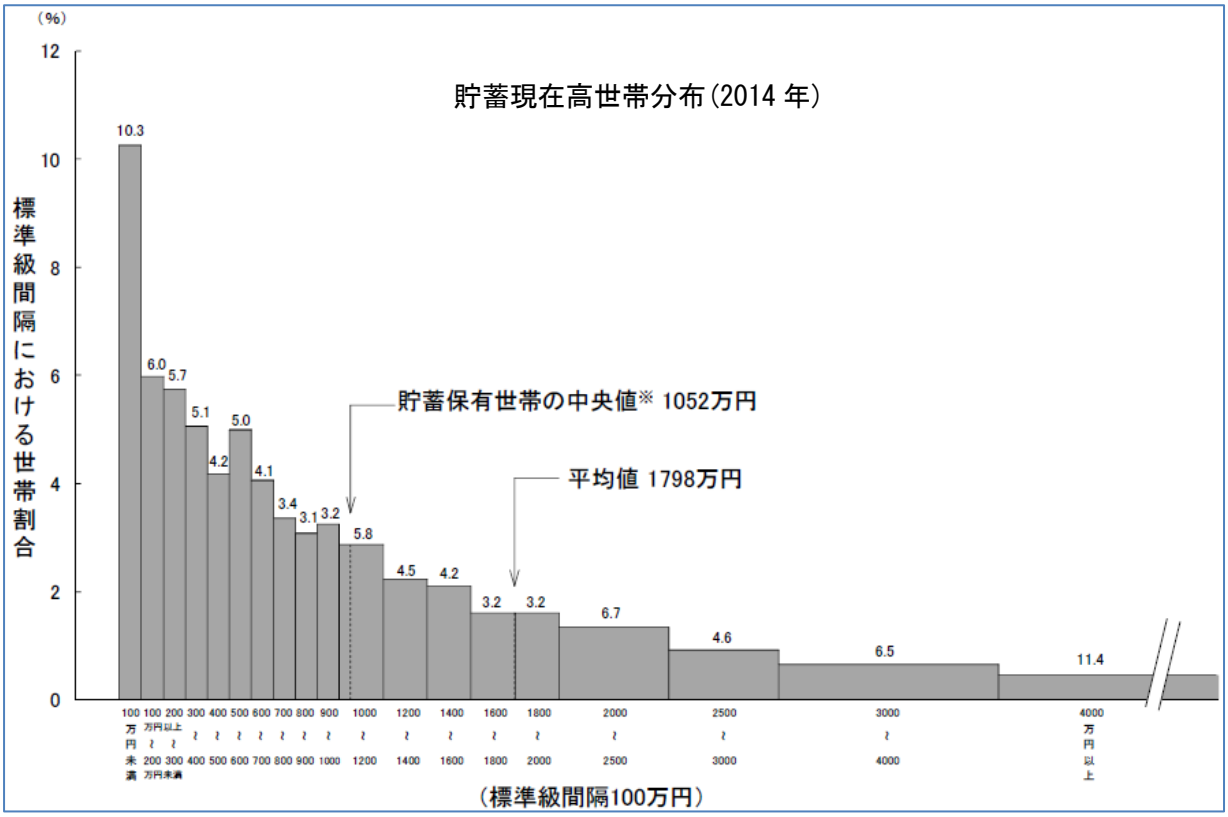
154.7　155.2　156.4　157.0　157.3　158.4　158.4　159.7　160.4　162.3

これより，中央値は $\frac{157.3+158.4}{2}=157.85$ (cm)

最頻値は，158.4 (cm)

column（メジアンとモード）

以下のグラフは平成 26 年の家計調査報告の「貯蓄現在高世帯分布」です。この分布によると貯蓄現在高の平均値は 1798 万円となっており，我々の実感と少しかけ離れています。平均値はきれいな山型の分布の場合は比較的私たちの実感と合いますが，偏った分布では少しずれが生じてしまいます。このようなときは中央値の 1052 万円の方が実感にあった値となるわけです。



また，ある電気店にパソコンを買いに行ったときに，店員から

「先月の各社の売り上げ台数の平均は 54 台でした」

と言われたらどう思いますか？ そんなことより，どちらかというと知りたいのは，

今，最も人気のあるパソコンは何なのか？

ではないだろうか？ つまりここで注目する数値は平均値ではなく最頻値なのである。

1ヵ月のパソコンの売り上げ台数（台）	
A社	36
B社	18
C社	72
D社	90
E社	63
F社	45
平均	54

← 最も人気のパソコン これを知りたい!!

例題1　次のデータは，A 班 5 人，B 班 6 人の 10 点満点のテストの結果である。

A 班：5，7，8，4，9　　B 班：7，10，9，4，8，6　（単位は点）

(1)　A 班のデータの平均値と B 班のデータの平均値をそれぞれ求めなさい。ただし，小数第 2 位を四捨五入しなさい。

(2)　A 班と B 班を合わせた 11 人のデータの平均値を求めなさい。

(3)　A 班のデータの中央値と B 班のデータの中央値をそれぞれ求めなさい。

練習1　次のデータは 10 人の生徒の 20 点満点のテストの結果である。

6，5，20，11，9，8，15，12，7，17　（単位は点）

(1)　このデータの平均値を求めなさい。

(2)　このデータの中央値を求めなさい。

例題2　右の表は，あるクラス 10 人について行われた数学のテストの得点の度数分布表である。得点はすべて整数とする。

(1)　このデータの平均値のとりうる値の範囲を求めなさい。

(2)　10 人の得点の平均点は 54.3 点であり，各得点は

69，65，62，57，55，55，48，42，x　（単位は点）

であった。x の値を求めなさい。

得点の階級(点)	人数
30 以上 40 未	1
40 ～ 50	2
50 ～ 60	4
60 ～ 70	3
計	10

練習2　右の表は，8 人の生徒について行われたテストの得点の度数分布表である。得点はすべて整数とする。

(1)　このデータの平均値のとりうる値の範囲を求めなさい。

(2)　8 人の得点の平均点は 52 点であり，各得点は

34，42，43，46，57，58，65，x　（単位は点）

であった。x の値を求めなさい。

得点の階級(点)	人数
20 以上 30 未	1
40 ～ 60	5
60 ～ 80	2
計	8

例題3　学生 9 人を対象に試験を行った結果，それぞれ 50，57，60，42，x，73，80，35，68 点だった。0 以上 100 以下の整数 x の値がわからないとき，このデータの中央値として何通りの値がありうるか。

練習3　次のデータは 10 人の生徒のある教科のテストの得点である。ただし，x の値は正の整数である。

43，55，x，64，36，48，46，71，65，50　（単位は点）

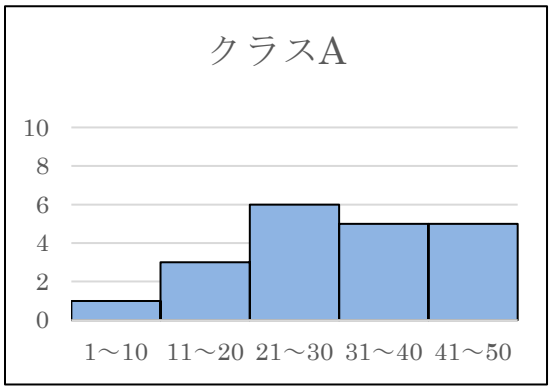
x の値が分からないとき，このデータの中央値として何通りの値がありうるか。

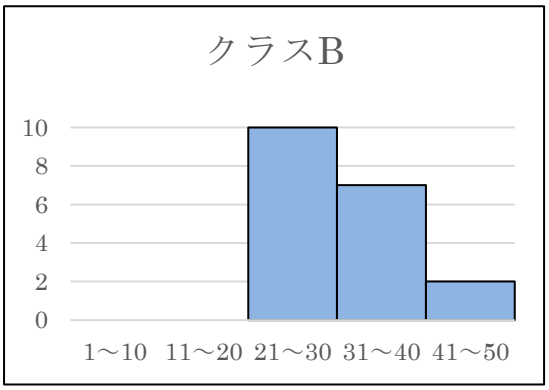
§2　四分位範囲

以下のデータは，2 つのクラス A，B に 50 点満点の数学のテストを行い，その結果を小さい方から順にならべたものである。平均値と中央値は 2 クラスとも同じ値となり，平均値は 30 点，中央値は 29 点であった。

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱	⑲	⑳
A	9	14	16	19	22	24	24	26	27	28	30	31	32	34	36	40	43	46	49	50
B	21	21	22	23	23	25	25	26	28	29	30	30	33	34	39	39	39	41	42	

2 つの代表値を見ている限りでは，2 つのクラスの得点分布に違いは見当たらない。しかし，ヒストグラムで表すと，その違いは一目瞭然である。クラス A の得点は高得点から，低得点まで万遍なく散らばっているのに対し，クラス B の得点は平均値付近に固まっていることが分かる。





ここでは，2 つのデータの**散らばり度合**を表す量について考えていく。

○範囲

データの最大値と最小値の差を**範囲**という。範囲の大きさは散らばり度合を表す 1 つの指標となる。

クラス A の得点の範囲は，$50-9=41$ 点

クラス B の得点の範囲は，$42-21=21$ 点

となるので，範囲を比べるとクラス A の得点の方が，散らばり度合が大きいと考えられる。

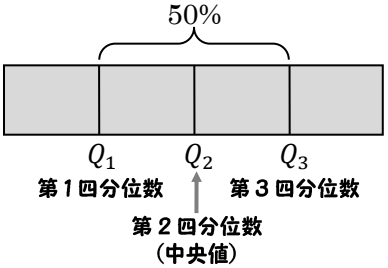
○四分位範囲

範囲は散らばり度合を比較する 1 つの方法であるが，最大値と最小値のみで決まるので，**極端な値の影響を受けやすい**という問題がある。

この問題を解決するために，データを大きさの順に並べて 4 等分し，中央の 50%の範囲を考えることがある。これを**四分位範囲**という。

データを 4 等分したとき，境目にある値を**四分位数**といい，小さい方から順に，**第 1 四分位数**，**第 2 四分位数**，**第 3 四分位数**という。記号では順に Q_1，Q_2，Q_3 と表す。なお，第 2 四分位数は中央値のことである。

これを用いると，四分位範囲は Q_3-Q_1 で求められる。



では，さっそく 2 クラスの得点データを用いて，四分位範囲を求めていこう。
まずは，四分位数から求めていく。

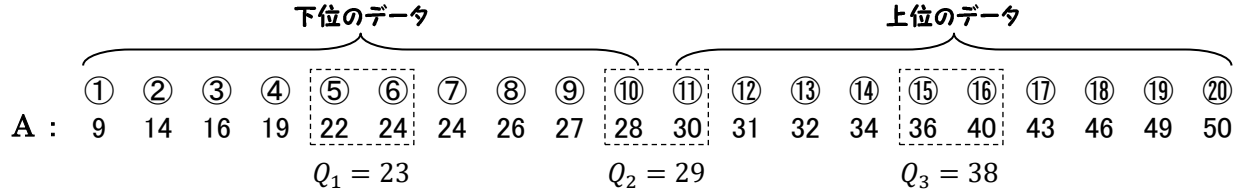
① クラス A の場合（データ数が偶数の場合）

まず，第 2 四分位数(中央値)は，小さい順に並べたとき，⑩と⑪の平均値となる。つまり，

$$Q_2 = \frac{28+30}{2} = 29 \text{ 点}$$

①～⑩のデータを下位のデータ，⑪～⑳のデータを上位のデータと呼ぶことにすると，
第 1 四分位数は下位のデータの中央値，第 3 四分位数は上位のデータの中央値となる。つまり，

$$Q_1 = \frac{22+24}{2} = 23 \text{ 点}, \quad Q_3 = \frac{36+40}{2} = 38 \text{ 点}$$



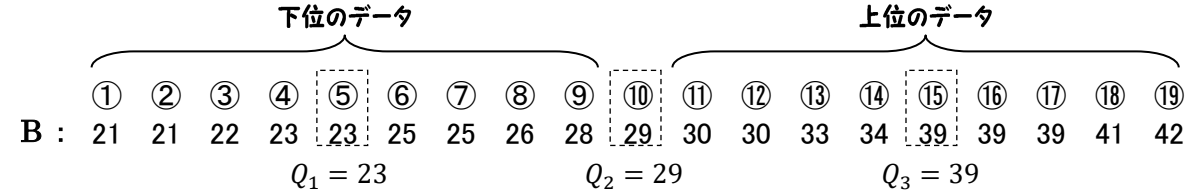
② クラス B の場合（データ数が奇数の場合）

まず，第 2 四分位数(中央値)は，小さい順に並べたとき，⑩の値となる。つまり，

$$Q_2 = 29 \text{ 点}$$

①～⑨のデータを下位のデータ，⑪～⑲のデータを上位のデータと呼ぶことにすると，
第 1 四分位数は下位のデータの中央値，第 3 四分位数は上位のデータの中央値となる。つまり，

$$Q_1 = \frac{22+24}{2} = 23 \text{ 点}, \quad Q_3 = \frac{36+40}{2} = 38 \text{ 点}$$



このように四分位数を求めるときは，**上位のデータと下位のデータの中央値**を考えればよい。
なお，データ数が奇数のときと偶数のときで，上位のデータと下位のデータのとり方が違うので注意が必要である。

以上のことから，

クラス A の得点の四分位範囲は，$Q_3 - Q_1 = 38 - 23 = 15$ 点

クラス B の得点の四分位範囲は，$Q_3 - Q_1 = 39 - 23 = 16$ 点

となるので，わずかにクラス B の得点の方が，散らばり度合が大きいと考えられる。

また，四分位範囲を 2 で割った値を**四分位偏差**といい，四分位範囲と同様に散らばり度合を比較する量の 1 つである。

○箱ひげ図

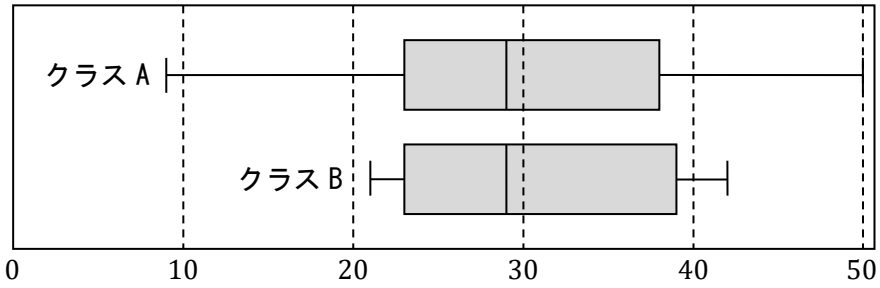
データのばらつきを比較するために，**箱ひげ図**と呼ばれる図を用いることがある。

箱ひげ図は，データの

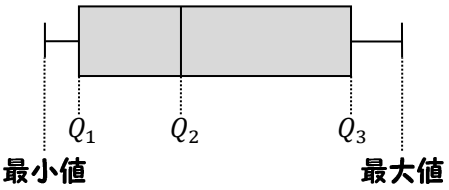
最小値，第 1 四分位数，第 2 四分位数(中央値)，第 3 四分位数，最大値

を箱と線(ひげ)を用いて表す図であり，先ほどのテストのデータを用いると，下図のようになる。

	最小値	第1四分位数	中央値	第3四分位数	最大値
クラスA	9	23	29	38	50
クラスB	21	23	29	39	42



ひげの左端が最小値，右端が最大値を表し，箱の左端が第 1 四分位数，右端が第 3 四分位数で，箱の中の仕切り線が第 2 四分位数を表している。



この箱ひげ図の全体を見ると，クラス A の方がばらつきが大きいと読み取ることができ，『箱』の部分をみると，わずかではあるがクラス B の方がばらつきが大きいと読み取れる。

箱ひげ図には多くの統計量が含まれており，複数のデータを比較するときには大変便利である。

例題4　次のデータは，A 班 10 人と B 班 9 人の 7 日間の勉強時間の合計を調べたものである。

A 班　5，15，17，11，18，22，12，9，14，4

B 班　2，16，13，19，6，3，10，8，7　(単位は時間)

(1)　それぞれのデータの範囲を求め，それに基づいて，データの散らばり度合いを比較しなさい。

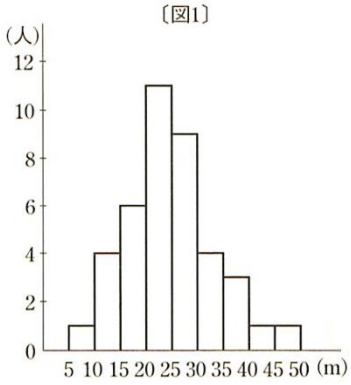
(2)　それぞれのデータの第 1 四分位数 Q_1，第 2 四分位数 Q_2，第 3 四分位数 Q_3 を求めなさい。

(3)　それぞれのデータの四分位範囲，四分位偏差を求めなさい。また，四分位範囲に基づいて，データの散らばり度合いを比較しなさい。

練習④　上の例題の A 班，B 班を合わせた大きさ 19 のデータの範囲，四分位偏差を求めなさい。

例題5　ある高校 3 年生 1 クラスの生徒 40 人について，ハンドボール投げの飛距離のデータを取った。〔図 1〕は，このクラスで最初に取ったデータのヒストグラムである。

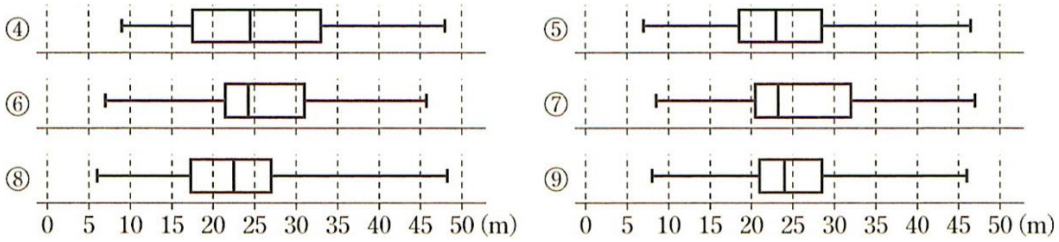
〔図1〕



(1)　この 40 人のデータの第 3 四分位数が含まれる階級を次の①～③から 1 つ選びなさい。

①　20m 以上 25m 未満　　②　25m 以上 30m 未満　　③　30m 以上 35m 未満

(2)　このデータを箱ひげ図にまとめたとき，〔図 1〕のヒストグラムと矛盾するものを次の④～⑨から 4 つ選びなさい。



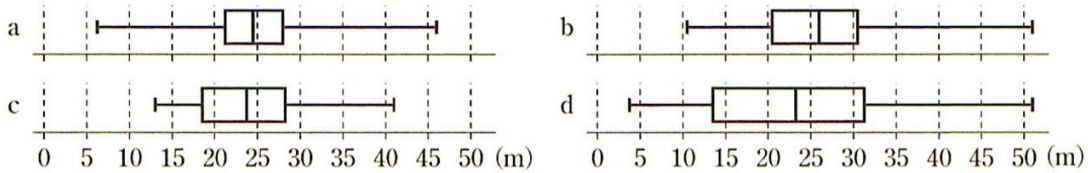
(3)　後日，このクラスでハンドボール投げの記録を取り直した。次に示した A～D は，最初に取った記録から今回の記録への変化の分析結果を記述したものである。a～d の各々が今回取り直したデータの箱ひげ図となる場合に，⑩～⑬の組合せのうち分析結果と箱ひげ図が矛盾するものを 2 つ選びなさい。

⑩　A-a　　⑪　B-b　　⑫　C-c　　⑬　D-d

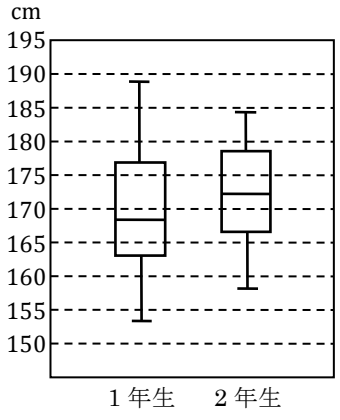
A：どの生徒の記録も下がった。　　B：どの生徒の記録も伸びた。

C：最初に取ったデータで上位 $\frac{1}{3}$ に入るすべての生徒の記録が伸びた。

D：最初に取ったデータで上位 $\frac{1}{3}$ に入るすべての生徒の記録は伸び，下位 $\frac{1}{3}$ に入るすべての生徒の記録は下がった。



練習5　右の図は，ある学校の 1 年生，2 年生各 200 人の身長のデータの箱ひげ図である。この箱ひげ図から読み取れることとして，正しいものを次の①～⑤からすべて選びなさい。



①　185 ㎝より大きい生徒が 1 年生にはいるが，2 年生にはいない。

②　170 ㎝以上の生徒が 1 年生では 100 人以下であるが，2 年生では 100 人以上いる。

③　165 ㎝以下の生徒がどちらの学年にも 50 人より多くいる。

④　175 ㎝以下の生徒は 1 年生では 150 人より多くいるが，2 年生では 150 人以下である。

⑤　1 年生のデータの範囲は，2 年生のデータの範囲より 20 ㎝以上大きい。

§3　分散・標準偏差

ここでは，**平均値からの散らばり度合**を表す量について考えていく。

ここでも，§2 で用いた 2 クラスのテストの結果のデータを用いて考えていこう。

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱	⑲	⑳
A	9	14	16	19	22	24	24	26	27	28	30	31	32	34	36	40	43	46	49	50
B	21	21	22	23	23	25	25	26	28	29	30	30	33	34	39	39	39	41	42	

今回は「平均値からどの程度離れているか？」を知りたいので，まずは平均値 30 と各データの差を考えていく。この値を**偏差**という。各データから 30 を引くと以下のようになる。

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱	⑲	⑳
A	−21	−16	−14	−11	−8	−6	−6	−4	−3	−2	0	1	2	4	6	10	13	16	19	20
B	−9	−9	−8	−7	−7	−5	−5	−4	−2	−1	0	0	3	4	9	9	9	11	12	

ここでは，各データが「平均して，どの程度平均値から離れているか？」を調べればよいが，偏差の平均値を計算すると，仮平均のところでも学んだとおり，正負が打ち消されて 0 になってしまう。
そこで，正負の影響をなくすために，**偏差の 2 乗の平均値**を考える。この値を**分散**という。

つまり，クラス A の分散は，

$$\frac{1}{20}\{(-21)^2+(-16)^2+(-14)^2+(-11)^2+(-8)^2+(-6)^2+(-6)^2+(-4)^2+(-3)^2+(-2)^2+$$
$$0^2+1^2+2^2+4^2+6^2+10^2+13^2+16^2+19^2+20^2\}$$
$$=\frac{2522}{20}=125$$

クラス B の分散は，

$$\frac{1}{19}\{(-9)^2+(-9)^2+(-8)^2+(-7)^2+(-7)^2+(-5)^2+(-5)^2+(-4)^2+(-2)^2+(-1)^2+$$
$$0^2+0^2+3^2+4^2+9^2+9^2+9^2+11^2+12^2\}$$
$$=\frac{928}{19}=48.842\cdots$$

となるので，クラス A の方が平均値からの散らばり具合が大きいことが分かる。

なお，分散は「偏差の **2 乗**の平均」なので単位をつけるとすれば，『点 ²』となる。
そこで，測定値と単位を揃えるために，**分散の平方根**を用いることがある。この値を**標準偏差**という。

つまり，クラス A の標準偏差は，$\sqrt{125}=11.180\cdots$（点）

クラス B の標準偏差は，$\sqrt{\frac{928}{19}}=6.988\cdots$（点）

となる。

一般的に，n 個のデータ x_1，x_2，x_3，…，x_{n-1}，x_n の平均値を $\bar{x}$，標準偏差を s と表すと，分散 s^2 は，次の式で表すことができる。

$$s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\ \cdots\ +(x_n-\bar{x})^2}{n}$$

$$= \frac{({x_1}^2-2x_1\bar{x}+\bar{x}^2)+({x_2}^2-2x_2\bar{x}+\bar{x}^2)+({x_3}^2-2x_3\bar{x}+\bar{x}^2)+\ \cdots\ +({x_n}^2-2x_n\bar{x}+\bar{x}^2)}{n}$$

$$= \frac{{x_1}^2+{x_2}^2+{x_3}^2+\ \cdots\ +{x_n}^2-2\bar{x}(x_1+x_2+x_3+\ \cdots\ +x_n)+n\bar{x}^2}{n}$$

$$= \frac{{x_1}^2+{x_2}^2+{x_3}^2+\ \cdots\ +{x_n}^2}{n}-2\bar{x}\cdot\frac{x_1+x_2+x_3+\ \cdots\ +x_n}{n}+\bar{x}^2$$

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

$$= \overline{x^2}-\bar{x}^2 \quad \left(\overline{x^2}\text{ はデータの各値の 2 乗の平均値を表す。}\right)$$

以上のことから，分散は (2 乗の平均値) − (平均値の 2 乗) で求めることもできる。

例 3　以下のテストの得点のデータにおいて，分散 s^2，標準偏差 s を求めなさい。

23　24　25　28　31　33　38　43　47　48　　(単位は点)

解①

平均値　$\dfrac{23+24+25+28+31+33+38+43+47+48}{10}=\dfrac{340}{10}=34$ (点)

よって，分散は

$$s^2=\frac{(-11)^2+(-10)^2+(-9)^2+(-6)^2+(-3)^2+(-1)^2+4^2+9^2+13^2+14^2}{10}=\frac{810}{10}=81$$

これより，標準偏差は，$s=\sqrt{81}=9$ (点)

なお，分散を求める際には，(2 乗の平均値) − (平均値の 2 乗) を用いてもよい。

解②

平均値　$\dfrac{23+24+25+28+31+33+38+43+47+48}{10}=\dfrac{340}{10}=34$ (点)

2 乗の平均値　$\dfrac{23^2+24^2+25^2+28^2+31^2+33^2+38^2+43^2+47^2+48^2}{10}=\dfrac{12370}{10}=1237$

よって，分散は　$s^2=1237-34^2=1237-1156=81$

column（標準偏差のとらえ方）

皆さんは平均点が 60 点のテストで 75 点を取ることができたらどのように思いますか？
「平均より，15 点も多く取れた！」と喜びますか？
それとも，「もっと高い点数を取らないとダメだ！」と自分を戒めますか？
この判断を正確にするための指標の 1 つが実は標準偏差になります。

例えば，このテストの標準偏差が 15 点だったとします。これは，「平均点からの離れ具合を平均値」が 15 点ということなので，$60-15=45$ (点)から $60+15=75$ (点)ぐらいまでの点数は，平均から標準的な離れ方ということになります。つまり，75 点というのは「割とよくある点数」ということなので，喜んでもいいですが，大騒ぎするほどいい点数ではないというわけです。

また，仮に標準偏差が 8 点だったとすると，75 点という点数は標準偏差の約 2 個分も離れていますから，かなり特殊なデータになります。つまり，この場合は大喜びする資格があるというわけです。

例題6　次のデータは，ある商品 A，B の 5 日間の売り上げ個数である。

A　5，7，4，3，6　　B　4，6，8，3，9　(単位は個)

A，B の変量をそれぞれ x，y とするとき，次の問いに答えなさい。

(1)　x，y のデータの平均値，分散，標準偏差をそれぞれ求めなさい。ただし，標準偏差については小数第 2 位を四捨五入しなさい。

(2)　x，y のデータについて，標準偏差によってデータの平均値からの散らばりの度合いを比較しなさい。

練習6　右の表は，A 工場，B 工場の同じ企画の製品 30 個の重さを測った結果である。

(1)　両工場のデータについて，平均値，標準偏差をそれぞれ求めなさい。ただし，小数第 3 位を四捨五入しなさい。

(2)　両工場のデータについて，標準偏差によってデータの平均値からの散らばりの度合いを比較しなさい。

製品の重さ(g)	個数	
	A 工場	B 工場
3.6	3	0
3.7	4	1
3.8	6	2
3.9	0	6
4.0	11	8
4.1	6	13
計	30	30

例題7　ある集団は A と B の 2 つのグループで構成されている。データを集計したところ，それぞれのグループの個数，平均値，分散は右の表のようになった。このとき，集団全体の平均値と分散を求めなさい。

グループ	個数	平均値	分散
A	20	16	24
B	60	12	28

練習7　12 個のデータがある。そのうちの 6 個のデータの平均値は 4，標準偏差は 3 であり，残りの 6 個のデータの平均値は 8，標準偏差は 5 である。

(1)　全体の平均値を求めなさい。

(2)　全体の分散を求めなさい。

例題8　次のデータは，ある都市のある年の月ごとの最高気温を並べたものである。

5，4，8，12，17，24，27，28，22，30，9，6　(単位は℃)

(1)　このデータの平均値を求めなさい。

(2)　このデータの中で入力ミスが見つかった。30℃となっている月の最高気温は正しくは 18℃であった。この入力ミスを修正すると，このデータの平均値は修正前より何℃減少しますか。

(3)　このデータの中で入力ミスが見つかった。正しくは 6℃が 10℃，30℃が 26℃であった。この入力ミスを修正すると，このデータの平均値は［　　］し，分散は［　　］する。

［　　］に当てはまるものを次の①，②，③から選びなさい。

①　修正前より増加　　②　修正前より減少　　③　修正前と一致

練習8　次のデータは，ある都市のある年の月ごとの最低気温を並べたものである。

−12，−9，−3，3，10，17，20，19，15，7，1，−8　(単位は℃)

(1)　このデータの平均値を求めなさい。

(2)　このデータの中で入力ミスが見つかった。正しくは−3℃が−1℃，3℃が 2℃，19℃が 18℃であった。この入力ミスを修正すると，このデータの平均値は［　　］し，分散は［　　］する。

［　　］に当てはまるものを上の例題の①，②，③から選びなさい。

○加工されたデータの平均値・標準偏差

ここでは，与えられたデータにある操作を行ったとき，平均値，標準偏差にどのような影響が出るのかを見ていく。

5 つのデータを 1，3，4，5，7 を用いて具体的に考えていこう。このデータの平均値，分散，標準偏差は以下のようになる。

平均値　$\dfrac{1+3+4+5+7}{5}=4$　　分散　$\dfrac{(-3)^2+(-1)^2+0^2+1^2+3^2}{5}=4$　　標準偏差　$\sqrt{4}=2$

データに定数を加える

上のデータの各値に 2 を加える。

1，3，4，5，7　⟹　3，5，6，7，9

このとき，平均値，分散，標準偏差は以下のようになる。

平均値 $\dfrac{3+5+6+7+9}{5}=6$　　分散 $\dfrac{(-3)^2+(-1)^2+0^2+1^2+3^2}{5}=4$　　標準偏差 $\sqrt{4}=2$

これを見て分かる通り，2 を加えることで，平均値が 2 増えているが，分散，標準偏差は変化していない。平均値が 2 増えるのは，データの各値が 2 ずつ増えているので当然と言える。そして，データの各値が 2 増え，平均値も 2 増えるので，偏差に変化は起こらない。このことから，分散，標準偏差も変化しないわけである。

一般的にデータ X の各値に a を加えたデータを Y とし，データ X の平均値，分散，標準偏差をそれぞれ $\bar{X}$，s_X^2，s_X，データ Y の平均値，分散，標準偏差をそれぞれ，$\bar{Y}$，s_Y^2，s_Y とすると，次の関係が成り立つ。

$$\bar{Y}=\bar{X}+a,\ s_Y^2=s_X^2,\ s_Y=s_X$$

データに定数を掛ける

上のデータの各値に 2 を掛ける。

$$1,\ 3,\ 4,\ 5,\ 7 \quad \Longrightarrow \quad 2,\ 6,\ 8,\ 10,\ 14$$

このとき，平均値，分散，標準偏差は以下のようになる。

平均値 $\dfrac{2+6+8+10+14}{5}=8$　　分散 $\dfrac{(-6)^2+(-2)^2+0^2+2^2+6^2}{5}=16$　　標準偏差 $\sqrt{16}=4$

これを見て分かる通り，2 を掛けることで，平均値が 2 倍，分散が 4 倍，標準偏差 2 倍に化している。平均値が 2 倍になるのは，データの各値が 2 倍になっているので当然と言える。そして，データの各値が 2 倍になり，平均値も 2 倍になるので，偏差も 2 倍になる。それを 2 乗して平均するので，分散は 2^2 倍，つまり 4 倍になる。標準偏差はそれの平方根なので 2 倍になるわけである。

一般的にデータ X の各値を a 倍したデータを Y とし，データ X の平均値，分散，標準偏差をそれぞれ $\bar{X}$，s_X^2，s_X，データ Y の平均値，分散，標準偏差をそれぞれ，$\bar{Y}$，s_Y^2，s_Y とすると，次の関係が成り立つ。

$$\bar{Y}=a\bar{X},\ s_Y^2=a^2s_X^2,\ s_Y=as_X$$

column（偏差値とは?）

テストを受けると点数や順位といったデータとともに，「偏差値」という値が算出されます。皆さんはこのデータの意味を知っていますか？ 「偏差値 40 はまずい…」「偏差値 60 なのでよく頑張った！」などのなんとなくのイメージを持っている人も多いと思いますが，ここではこの偏差値の意味を詳しく解説していきます。

偏差値というのは，平均値が 50，標準偏差が 10 となるようにデータの各値を変換したものになります。具体的に見ていきましょう。

n 個のデータ x_1，x_2，x_3，…，x_n の平均値を $\bar{x}$，標準偏差を s とします。

まずはデータの各値から平均値の $\bar{x}$ を引いて，平均値を 0 に変換します。

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

次にこれらを s で割り，標準偏差を 1 に変換します。

$$\frac{x_1-\bar{x}}{s},\ \frac{x_2-\bar{x}}{s},\ \frac{x_3-\bar{x}}{s},\ \cdots,\ \frac{x_n-\bar{x}}{s}$$

最後にこれらに 10 を掛け，50 を足すことで平均値 50，標準偏差 10 に変換します。

$$\frac{x_1-\bar{x}}{s}\times 10+50,\ \frac{x_2-\bar{x}}{s}\times 10+50,\ \frac{x_3-\bar{x}}{s}\times 10+50,\ \cdots,\ \frac{x_n-\bar{x}}{s}\times 10+50$$

この変換をすることで，偏差値が 40～60 であれば標準偏差 1 個分以内の離れ方なので標準的な点数であり，偏差値が 30 または 70 あたりになると，平均から標準偏差 2 個分離れているのでかなり特殊な点数であることが分かります。

また，ある試験の結果，数学の点数 80 点，国語の点数が 60 点であったとき，この点数だけで「数学の方が，結果が良かった！」と言えるでしょうか？ たとえ80 点でも平均点が 80 点であれば標準的な点数ということになりますし，60 点であってもそれが全体の最高得点ということもあり得ます。このようなときに偏差値を比べるわけです。もし，数学の偏差値が 55 で，国語の偏差値が 70 であれば，国語の結果の方がよかったという結論になるわけです。

また，数学のテストにおいて，1 学期が 80 点，2 学期が 60 点を取ったとき，やはりこの点数だけでは「成績が下がった…」ということにはなりません。当然，1 学期と 2 学期でテストの難易度に差があるかもしれないからです。ここで偏差値を比べ，1 学期の偏差値が 60 で，2 学期の偏差値が 65 になったというのであれば，成績が少し上がったということになります。

このように，偏差値を考えることで，科目の違いや難易度の違いを気にすることなく，同じ基準でテストの結果を評価することができるわけです。

例題9　変量 x のデータの平均値 $\bar{x}$ が $\bar{x}=21$，分散 ${s_x}^2$ が ${s_x}^2=12$ であるとする。このとき，次の式によって得られる新しい変量 y のデータについて，平均値 $\bar{y}$，分散 ${s_y}^2$，標準偏差 s_y を求めなさい。

ただし，$\sqrt{3}=1.73$ とし，標準偏差は小数第 2 位を四捨五入して，小数第 1 位まで求めなさい。

(1)　$y=x-5$　　(2)　$y=3x$　　(3)　$y=-2x+3$　　(4)　$y=\dfrac{x-21}{2\sqrt{3}}$

練習9　ある変量のデータがあり，その平均値は 50，標準偏差は 15 である。そのデータを修正して，各データの値を 1.2 倍して 5 をひいたとき，修正後の平均値と標準偏差を求めなさい。

例題10　次の変量 x のデータについて，以下の問いに答えなさい。

726，814，798，750，742，766，734，702

(1)　$y=x-750$ とおくことにより，変量 x のデータの平均値 $\bar{x}$ を求めなさい。

(2)　$u=\dfrac{x-750}{8}$ とおくことにより，変量 x のデータの分散を求めなさい。

練習10　次の変量 x のデータについて，以下の問いに答えなさい。

514，584，598，521，605，612，577

(1)　$y=x-570$ とおくことにより，変量 x のデータの平均値 $\bar{x}$ を求めなさい。

(2)　$u=\dfrac{x-570}{7}$ とおくことにより，変量 x のデータの分散を求めなさい。

§4　データの相関

ここでは，2 つの変量の関係を調べる方法について考えていく。

以下の表は，高校 1 年生 40 名の身長 x (cm)，体重 y (kg)の関係を調べたものである。

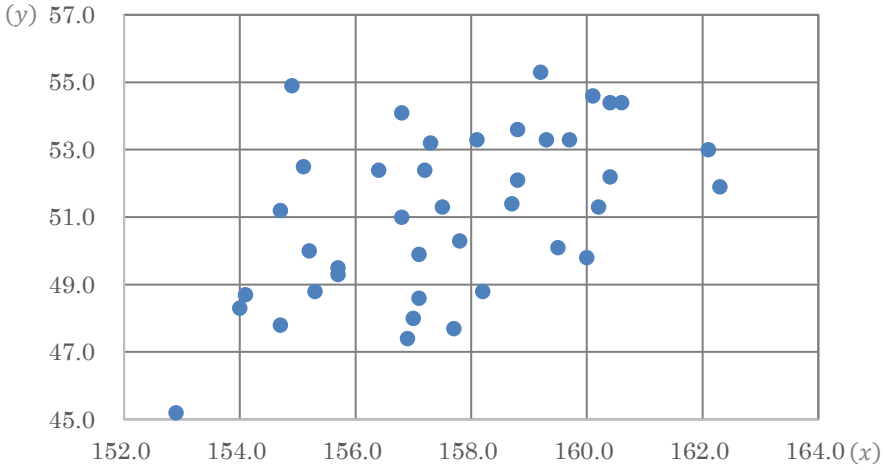
	x	y
①	152.9	45.2
②	154.0	48.3
③	154.1	48.7
④	154.7	47.8
⑤	155.3	48.8
⑥	155.2	50.0
⑦	154.7	51.2
⑧	155.1	52.5
⑨	154.9	54.9
⑩	155.7	49.3

	x	y
⑪	155.7	49.5
⑫	156.4	52.4
⑬	156.9	47.4
⑭	157.0	48.0
⑮	157.1	48.6
⑯	157.1	49.9
⑰	156.8	51.0
⑱	157.2	52.4
⑲	157.3	53.2
⑳	156.8	54.1

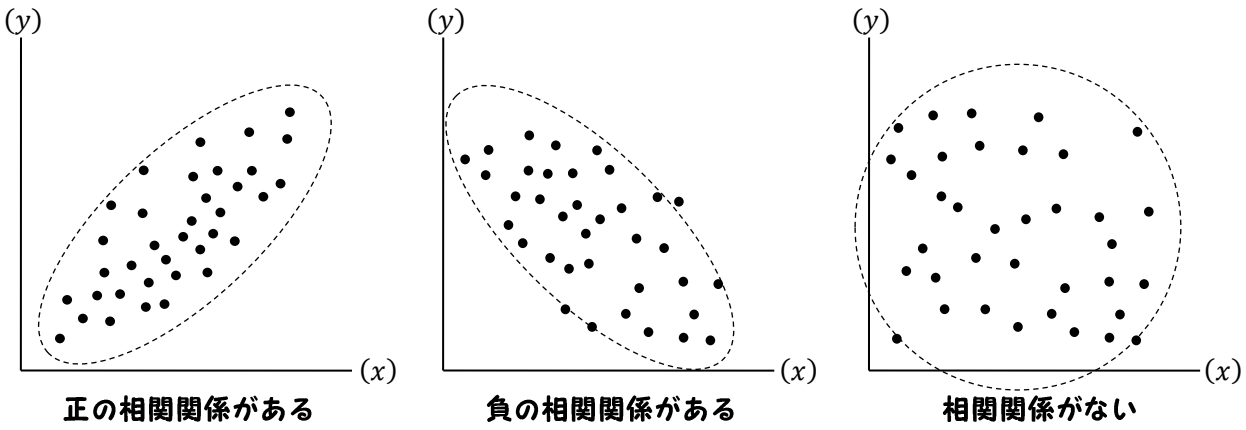
	x	y
㉑	157.7	47.7
㉒	158.2	48.8
㉓	157.8	50.3
㉔	157.5	51.3
㉕	158.1	53.3
㉖	159.5	50.1
㉗	158.7	51.4
㉘	158.8	52.1
㉙	159.3	53.3
㉚	158.8	53.6

	x	y
㉛	159.2	55.3
㉜	160.0	49.8
㉝	160.2	51.3
㉞	160.4	52.2
㉟	159.7	53.3
㊱	160.4	54.4
㊲	160.1	54.6
㊳	160.6	54.4
㊴	162.3	51.9
㊵	162.1	53.0

この x と y の関係を分かりやすくするために，x と y の組 $(x,\ y)$ を座標とする点を以下のような座標平面上にとっていく。このような図のことを**散布図**という。



この図から，身長が増えていくと体重も増えていくという傾向が読み取れる。このとき，2 つの変量 x，y の間に，**正の相関関係**があるという。逆に，片方が増えるともう片方は減る傾向にあるとき，2 つの変量は**負の相関関係**があるといい，どちらの傾向も認められないときは**相関関係がない**という。



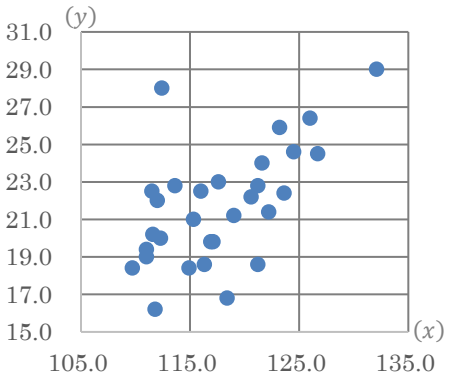
○相関係数

以下の表と図は，小学１年生の身長と体重の関係を表したものである。

	x	y
①	111.5	22.5
②	111.6	20.2
③	118.4	16.8
④	126.7	24.5
⑤	116.0	22.5
⑥	117.1	19.8
⑦	115.3	21.0
⑧	109.7	18.4
⑨	121.2	18.6
⑩	112.4	28.0

	x	y
⑪	111.0	19.0
⑫	114.9	18.4
⑬	116.3	18.6
⑭	121.2	22.8
⑮	120.6	22.2
⑯	112.0	22.0
⑰	122.2	21.4
⑱	111.8	16.2
⑲	111.0	19.4
⑳	132.1	29.0

	x	y
㉑	123.6	22.4
㉒	126.0	26.4
㉓	119.0	21.2
㉔	116.9	19.8
㉕	117.6	23.0
㉖	124.5	24.6
㉗	121.6	24.0
㉘	113.6	22.8
㉙	112.3	20.0
㉚	123.2	25.9



これも，先ほどの高校１年生の身長，体重と同様に正の相関関係があること分かるが，では，「どちらの相関関係がより強いのか？」と聞かれたらどうだろうか。関係の強さを図だけで判断するのはさすがに難しいので，ここでは相関関係の正負と強弱を数値で表す方法を考えていく。

2 つの変量 x，y のデータが，$(x,\ y)$ の値の組として次のように与えられたとする。

$$(x_1,\ y_1),\ (x_2,\ y_2),\ (x_3,\ y_3),\ \cdots,\ (x_n,\ y_n)$$

以下，$x_1,\ x_2,\ \cdots,\ x_n$ の平均値を $\bar{x}$，標準偏差を s_x とし，$y_1,\ y_2,\ \cdots,\ y_n$ の平均値を $\bar{y}$，標準偏差を s_y とする。このとき，まずは「x の偏差と y の偏差の積」の平均値を考える。

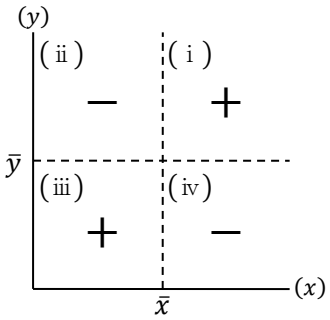
$$\frac{1}{n}\{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+(x_3-\bar{x})(y_3-\bar{y})+\ \cdots\ +(x_n-\bar{x})(y_n-\bar{y})\}$$

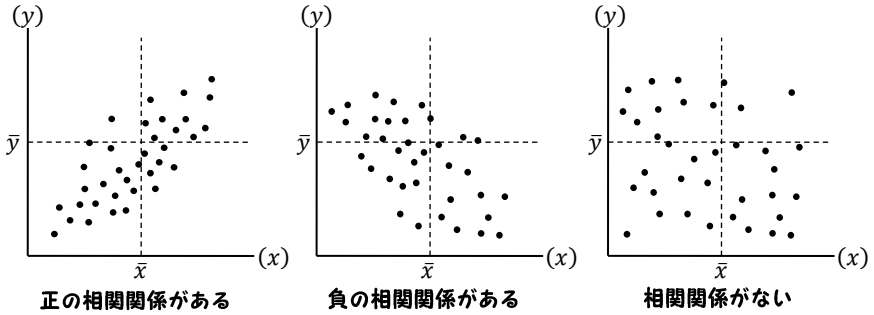
これを，x と y の**共分散**といい，s_{xy} と表す。

共分散の意味

偏差の積が正 $(x_k-\bar{x})(y_k-\bar{y})>0$ のとき，２つの偏差はともに正，またはともに負となるので，$(x_k,\ y_k)$ は右図の(ⅰ)または(ⅲ)の領域に含まれる。

また，偏差の積が負 $(x_k-\bar{x})(y_k-\bar{y})<0$ のとき，２つの偏差の正負が異なるので，$(x_k,\ y_k)$ は右図の(ⅱ)または(ⅳ)の領域に含まれる。





これより，正の相関関係があるとき，多くの点が(ⅰ)または(ⅲ)の領域に含まれるので，共分散の値は大きくなる。逆に負の相関関係があるとき，多くの点が(ⅱ)または(ⅳ)の領域に含まれるので，共分散の値は小さくなる。そして，相関関係がないときは領域内に万遍なく広がるので，共分散は 0 に近い値になる。

そして，この共分散を x と y の標準偏差の積 $s_x s_y$ で割ったものを**相関係数**といい，r で表す。

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

「標準偏差の積で割る」ことで，相関係数は必ず $-1 \leqq r \leqq 1$ の範囲に収まる。

先ほどの，高 1(40 名)と小 1(30 名)の共分散と標準偏差は次のようになる。

	高1	小1
共分散	2.87	9.89
相関係数	0.51	0.56

共分散は，与えられたデータの個数や大きさの影響を受けるので，異なるデータを比較するのに適していないが，相関係数は必ず $-1 \leqq r \leqq 1$ の範囲に収まるので，異なるデータの比較に適している。高 1 と小 1 のデータを比較すると，小 1 の身長・体重の方が，わずかに相関が強いことが分かる。

一般的に，相関係数と相関の強さの関係は以下のようになっていることが知られている。

相関係数	相関の強さ
$r = 0$	相関なし
$0 < \lvert r \rvert \leqq 0.2$	ほとんど相関なし
$0.2 < \lvert r \rvert \leqq 0.4$	弱い相関あり
$0.4 < \lvert r \rvert \leqq 0.7$	相関あり
$0.7 < \lvert r \rvert < 1$	強い相関あり
$\lvert r \rvert = 1$	完全な相関

例 4　以下のデータは生徒 10 人の化学と数学のテストの点数である。相関係数 r を求めなさい。

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
化学(x 点)	70	80	50	30	90	60	70	20	40	90
数学(y 点)	80	90	40	50	70	60	50	30	20	60

化学の平均点を $\bar{x}$，標準偏差を s_x，数学の平均点を $\bar{y}$，標準偏差を s_y とし，共分散を s_{xy} とする。
また，化学，数学の各点数を x_k，y_k　$(k = 1,\ 2,\ 3,\ \cdots,\ 10)$ とする。
表より，

$$s_x = \frac{1}{10}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ \cdots + (x_{10} - \bar{x})^2\} = \sqrt{540}$$

$$s_y = \frac{1}{10}\{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ \cdots + (y_{10} - \bar{y})^2\} = \sqrt{220}$$

$$s_{xy} = \frac{1}{10}\{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ \cdots + (x_{10} - \bar{x})(y_{10} - \bar{y})\} = 190$$

となるので，$r = \dfrac{190}{\sqrt{540} \cdot \sqrt{220}} = \dfrac{19}{6\sqrt{33}} = 0.55124\ \cdots$

	x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
①	70	60	10	100	10	100	100
②	80	30	20	400	−20	400	−400
③	50	40	−10	100	−10	100	100
④	30	40	−30	900	−10	100	300
⑤	90	80	30	900	30	900	900
⑥	60	50	0	0	0	0	0
⑦	70	50	10	100	0	0	0
⑧	20	30	−40	1600	−20	400	800
⑨	40	60	−20	400	10	100	−200
⑩	90	60	30	900	10	100	300
計	600	500	0	5400	0	2200	1900
平均値	60	50	0	540	0	220	190

（相関係数の式の意味）

相関係数の式は覚えづらい形をしていますが，少し見方を変えることで，とらえやすくなります。

2 つの n 個のデータ

$$x：x_1，\ x_2，\ x_3，\ \cdots，\ x_n \qquad y：y_1，\ y_2，\ y_3，\ \cdots，\ y_n$$

の平均値を $\bar{x}$，$\bar{y}$，標準偏差を s_x，s_y，共分散を s_{xy} とする。

このとき，$r=\dfrac{s_{xy}}{s_x s_y}$

$$=\frac{\dfrac{1}{n}\{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+(x_3-\bar{x})(y_3-\bar{y})+\ \cdots\ +(x_n-\bar{x})(y_n-\bar{y})\}}{s_x s_y}$$

$$=\frac{1}{n}\left(\frac{x_1-\bar{x}}{s_x}\cdot\frac{y_1-\bar{y}}{s_y}+\frac{x_2-\bar{x}}{s_x}\cdot\frac{y_2-\bar{y}}{s_y}+\frac{x_3-\bar{x}}{s_x}\cdot\frac{y_3-\bar{y}}{s_y}+\ \cdots\ +\frac{x_n-\bar{x}}{s_x}\cdot\frac{y_n-\bar{y}}{s_y}\right)$$

ここで，2 つの n 個のデータを X，Y とすると，

$$X：\frac{x_1-\bar{x}}{s_x}，\ \frac{x_2-\bar{x}}{s_x}，\ \frac{x_3-\bar{x}}{s_x}，\ \cdots，\ \frac{x_n-\bar{x}}{s_x} \qquad Y：\frac{y_1-\bar{y}}{s_y}，\ \frac{y_2-\bar{y}}{s_y}，\ \frac{y_3-\bar{y}}{s_y}，\ \cdots，\ \frac{y_n-\bar{y}}{s_y}$$

X，Y は x，y の各値から平均値を引き，標準偏差で割ったデータになります。

つまり，X，Y は x，y を**平均値 0,標準偏差 1 となるように加工したデータ**になります。

このようにすることで，異なるデータの比較ができるようになるわけです。

そして，相関係数というのは，平均値 0，標準偏差 1 となるように加工したデータに対して共分散をとっていると見ることができます。

例題11　次のような変量 x, y のデータがある。これらについて，散布図をかき，x と y の間に相関関係があるかどうか調べなさい。また，相関関係がある場合には，正・負のどちらであるかいいなさい。

(1)

x	1	3	8	5	4	6	2	9
y	2	2	6	7	3	5	3	8

(2)

x	38	46	20	48	18	27	11	33
y	12	15	25	11	30	21	38	30

(3)

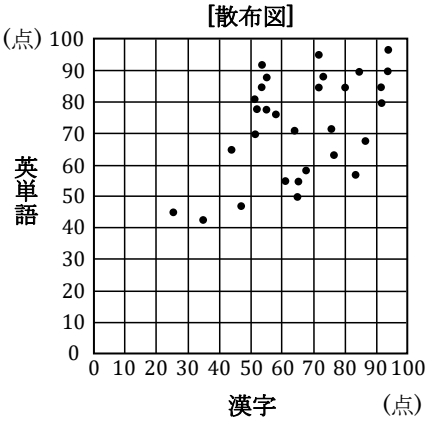
x	1.3	3.3	4.9	2.2	5.7	3.6	2.7	4.0
y	2.6	4.2	2.0	1.3	4.2	1.2	4.1	3.6

練習11　右の散布図は，30 人のクラスの漢字と英単語の 100 点満点で実施したテストの得点の散布図である。

(1)　この散布図をもとにして，漢字と英単語の得点の間に相関関係があるかどうか調べなさい。また，相関関係がある場合には正・負のどちらであるかをいいなさい。

(2)　この散布図をもとにして，英単語の度数分布表を作成しなさい。ただし，階級は[40 以上 50 未満]，…[90 以上 100 未満]とする。

[散布図]



例題12　次の表は，学生 5 名の身長 x (cm)と体重 y (kg)を測定した結果である。x と y の相関係数 r を求めなさい。

	A	B	C	D	E
身長 x (cm)	181	167	173	169	165
体重 y (kg)	75	59	63	67	61

練習12　下の表は，10 人の生徒に 30 点満点の 2 種類のテスト A，B を行った得点の結果である。テスト A，B の得点をそれぞれ x, y とするとき，x と y の相関係数 r を求めなさい。ただし，小数第 3 位を四捨五入しなさい。

生徒の番号	1	2	3	4	5	6	7	8	9	10
x	29	25	22	28	18	23	26	30	30	29
y	23	23	18	26	17	20	21	20	26	26

例題13　右の表は，10 名からなるある少人数クラスで，100 点満点で 2 回ずつ実施した数学と英語のテストの得点のまとめたものである。

(1)　数学と英語の得点の散布図を，1 回目，2 回目の各回についてかきなさい。

(2)　1 回目の数学と英語の得点の相関係数を r_1，2 回目の数学と英語の得点の相関係数を r_2 とするとき，値の組 $(r_1,\ r_2)$ として正しいものを以下の①～④から 1 つ選びなさい。

①　$(0.54,\ 0.20)$　　②　$(-0.54,\ 0.20)$

③　$(0.20,\ 0.54)$　　④　$(0.20,\ -0.54)$

	1 回目		2 回目	
番号	数学	英語	数学	英語
1	40	43	60	54
2	63	55	61	67
3	59	62	56	60
4	35	64	60	71
5	43	36	69	80
6	36	48	64	50
7	51	46	54	57
8	57	71	59	40
9	32	65	49	42
10	34	50	57	69

練習13　右の表は，2 つの変量 x，y のデータである。

(1)　これらのデータについて，$0.72,\ -0.19,\ -0.85$ のうち，x，y の相関係数に最も近いものはどれですか。

(2)　表の右端のデータの y の値を 68 に変更すると，x と y の相関係数の絶対値は大きくなるか，それとも小さくなるか。

x	80	70	62	72	90	78
y	58	72	83	71	52	78