データの分析

1 データの整理

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

(1) テレビをみた時間が 85 分未満の生徒は何人いるか。

(2) テレビをみた時間が 95 分以上の生徒は全体の何%であるか。

(3) 右の度数分布表をもとにして，ヒストグラムをかけ。

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

要 点

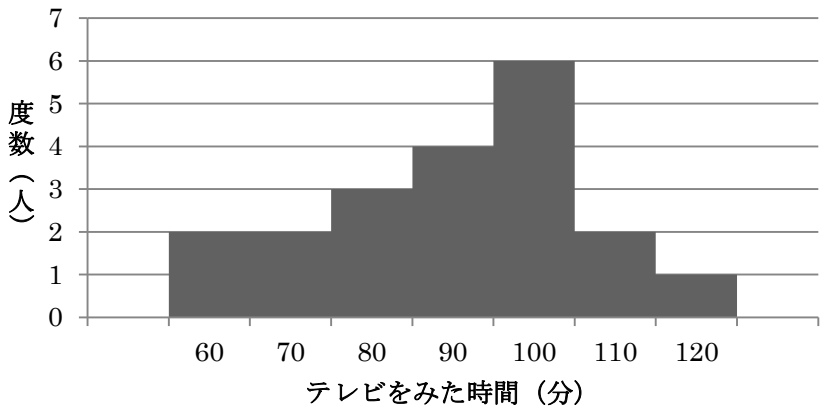
ある集団を構成する人や物の特性を表す数量を **変量** といい，変量の個々の値や，その集まりのことを **データ** という。
データを整理するとき，設定した各区間を **階級** といい，各階級の中央の値をその階級の **階級値** という。また，各階級に入る値の個数をその階級の **度数** といい，各階級に度数を対応させた表を **度数分布表** という。データ全体に対する各階級の度数の割合を，その階級の **相対度数** という。
縦軸に度数，横軸に階級をとったグラフを **ヒストグラム** という。

解答

(1) 2＋2＋3＝**7**（人）

(2) 0.30＋0.10＋0.05＝0.45
したがって **45%**

(3)



2　平均値

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

10，8，4，6，9，5，7 （回）

このデータの平均値 $\overline{x}$ を求めよ。

(2)　右の表から，テレビをみた時間 x の平均値を求めよ。

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

要　点

平均値

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

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

度数分布表からの平均値

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

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

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

$$\overline{x}=\frac{1}{n}(x_1f_1+x_2f_2+\cdots\cdots+x_rf_r)$$

ただし　$n=f_1+f_2+\cdots\cdots+f_r$

解答

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

(2)　$\overline{x}=\frac{1}{20}(60\times2+70\times2+80\times3+90\times4+100\times6+110\times2+120\times1)$

$=\frac{1}{20}\times1800=$**90（分）**

3　中央値，最頻値

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

3，2，0，1，3，1，1，2 （冊）

(1)　このデータの中央値を求めよ。

(2)　このデータの最頻値を求めよ。

要　点

中央値

データの値を大きさの順に並べたとき，中央の順位にくる値を **中央値** または **メジアン** という。

データの値の個数が偶数のときは，中央に並ぶ 2 つの値の平均値を中央値とする。

最頻値

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

解答

(1)　小さい方から順に並べると　0，1，1，1，2，2，3，3

これより，中央値は　$\frac{1+2}{2}$ ＝**1.5（冊）**

(2)　最頻値は　**1（冊）**

4　範囲，四分位数，四分位範囲，四分位偏差

次のデータは，A 社の従業員 11 人の年収を調べたものである。

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

次の問いに答えよ。

(1)　このデータの範囲を求めよ。

(2)　このデータの四分位数 Q_1，Q_2，Q_3 を求めよ。

(3)　このデータの四分位範囲と四分位偏差を求めよ。

要　点

範囲

データの最大値から最小値を引いた値を **範囲** という。

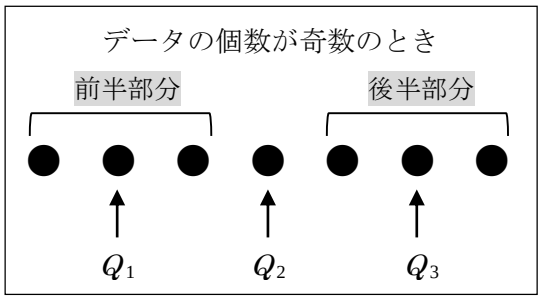
範囲 ＝ 最大値 － 最小値

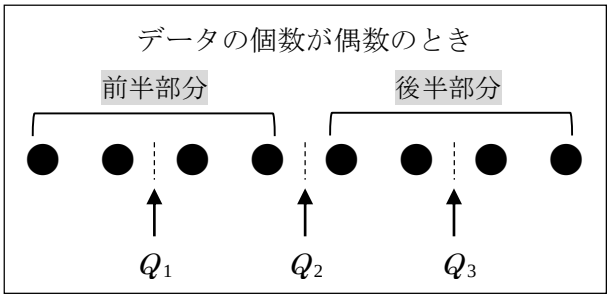
四分位数

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

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

これらをまとめて **四分位数** という。





四分位範囲，四分位偏差

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

また，四分位範囲を 2 で割った値を **四分位偏差** という。

四分位範囲 ＝ $\boldsymbol{Q_3-Q_1}$ ，　　四分位偏差＝$\dfrac{\boldsymbol{Q_3-Q_1}}{\boldsymbol{2}}$

解答

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

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

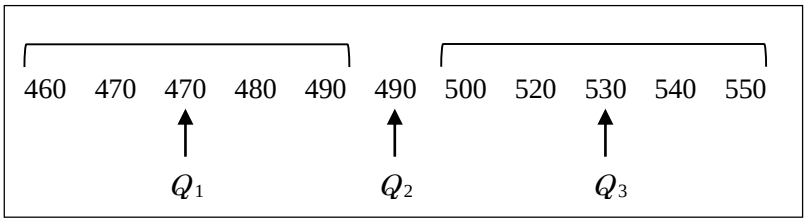
中央値から　**Q_2＝490（万円）**

前半部分の中央値から

Q_1＝470（万円）

後半部分の中央値から

Q_3＝530（万円）



(3)　Q_1＝470，Q_3＝530 であるから

四分位範囲は　530－470＝**60（万円）**

四分位偏差は　$\dfrac{60}{2}$ ＝**30（万円）**

5 **箱ひげ図**

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

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

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

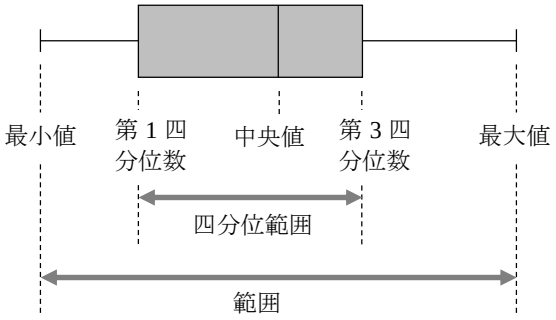
要 点

箱ひげ図

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

箱ひげ図から，範囲や四分位範囲を読み取ることもできる。

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



解答

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

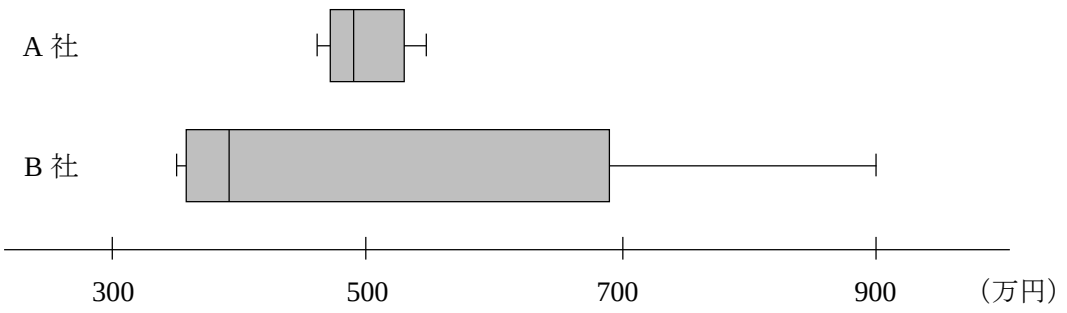
B 社の 最小値，Q_1，Q_2，Q_3，最大値を求める。

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

これから，最小値，最大値は 350，900 （万円）

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

以上から，A 社と B 社の箱ひげ図は次のようになる。



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

6　分散

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

10，8，4，6，9，5，7 （回）

このデータの分散 s^2 を求めよ。

要　点

分散

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

$$\textbf{分散＝（偏差）}^2\textbf{ の平均値 ，}\qquad s^2=\frac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots\cdots+(x_n-\overline{x})^2\}$$

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

$$\textbf{分散＝（}x^2\textbf{ の平均値）－（}x\textbf{ の平均値）}^2\textbf{ ，}\qquad s^2=\overline{x^2}-\left(\overline{x}\right)^2$$

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

$$s^2=\frac{1}{n}\{(x_1-m)^2+(x_2-m)^2+\cdots\cdots+(x_n-m)^2\}$$

$$=\frac{1}{n}\{(x_1^2-2x_1m+m^2)+(x_2^2-2x_2m+m^2)+\cdots\cdots+(x_n^2-2x_nm+m^2)\}$$

$$=\frac{1}{n}(x_1^2+x_2^2+\cdots\cdots+x_n^2-2x_1m-2x_2m-\cdots\cdots-2x_nm+m^2+m^2+\cdots\cdots+m^2)$$

$$=\frac{1}{n}\{(x_1^2+x_2^2+\cdots\cdots+x_n^2)-2m(x_1+x_2+\cdots\cdots+x_n)+n\cdot m^2\}$$

$$=\frac{1}{n}(x_1^2+x_2^2+\cdots\cdots+x_n^2)-2m\cdot\frac{1}{n}(x_1+x_2+\cdots\cdots+x_n)+m^2$$

$$=\frac{1}{n}(x_1^2+x_2^2+\cdots\cdots+x_n^2)-2m\cdot m+m^2=\frac{1}{n}(x_1^2+x_2^2+\cdots\cdots+x_n^2)-m^2$$

解答

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

偏差は　3，1，−3，−1，2，−2，0 （回）

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

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

したがって　$s^2=53-7^2=4$

7　標準偏差

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

4，2，1，1，4，2，0 （冊）

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

要　点

標準偏差

分散の正の平方根を **標準偏差** といい，s で表す。

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

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

標準偏差＝$\sqrt{(x^2\text{の平均値})-(x\text{の平均値})^2}$ ，　　$s=\sqrt{\overline{x^2}-\left(\overline{x}\right)^2}$

解答

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

偏差は　2，0，−1，−1，2，0，−2 （冊）

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

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

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

8　散布図

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

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

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

要　点

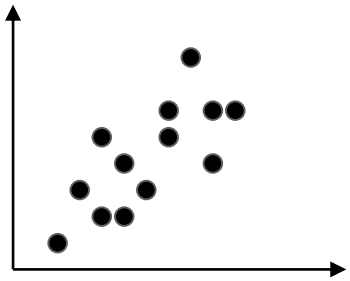
散布図

2 つの変量の値の組を座標平面上の点で表したものを **散布図** という。

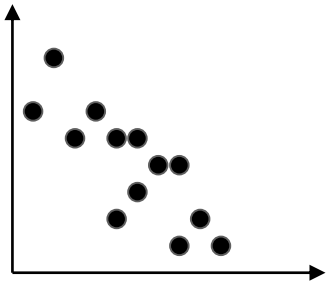
散布図と相関関係

2 つの変量 x，y について，

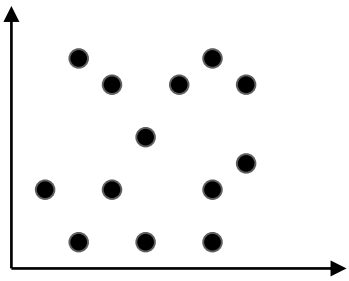
一方の値が大きくなると他方の値も大きくなる傾向があるとき，x と y の間には **正の相関関係がある** という。



一方の値が大きくなると他方の値は小さくなる傾向があるとき，x と y の間には **負の相関関係がある** という。

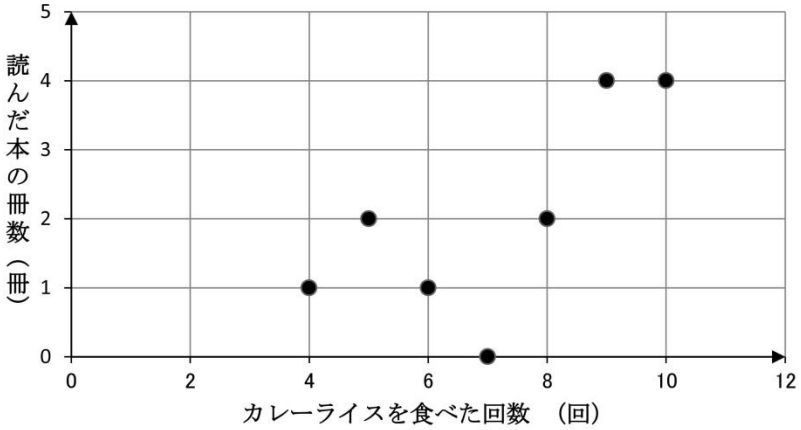


正，負いずれの相関関係も見られないとき，x と y の間には **相関関係がない** という。



解答

散布図は右のようになる。



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

9　相関係数

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

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

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

要　点

共分散

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

共分散 ＝ 偏差の積の平均値

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

相関係数

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

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

分母と分子に n を掛けると，次の式が得られる。

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

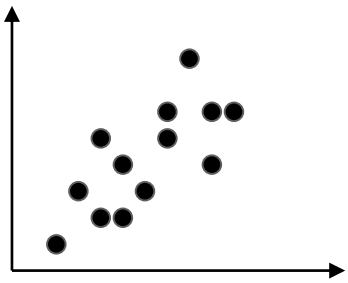
$$r=\frac{(x_1-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{\sqrt{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}\sqrt{(y_1-\overline{y})^2+\cdots+(y_n-\overline{y})^2}}$$

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

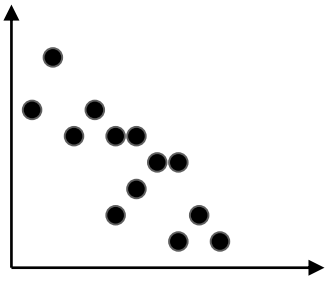
r が正のとき，正の相関関係がある。
r が 1 に近い値であるほど，正の相関関係が強い。

r が負のとき，負の相関関係がある。
r が -1 に近い値であるほど，負の相関関係が強い。

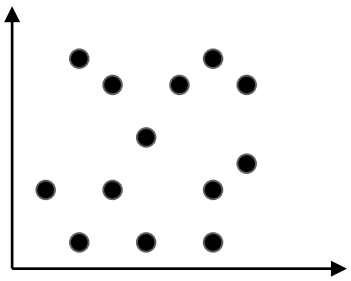
r が 0 に近い値であるほど，相関関係が弱い。



$r \fallingdotseq 0.73$



$r \fallingdotseq -0.81$



$r \fallingdotseq 0.14$

解答

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

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

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

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