第 II 部　統計

6　共分散と相関係数

ポイント

- 変量間の関係を表す指標
- 平均からの偏差
- 共分散
- 相関係数

6.1　はじめに

- マーケティングミックスは需要に影響を与えると仮定しています。
- マーケティングミックスの項目と需要の関係を数字であらわす方法を考えます。

6.1.1　四則演算の確認

展開の公式を示す。

$$(a-b)\,c = ac - bc \tag{1}$$

$$(a-b)\,(c-d) = ac - ad - bc + bd \tag{2}$$

$$(a-b)^2 = a^2 - 2ab + b^2 \tag{3}$$

(1) から (2) の導出

(1) 式において $c = u - v$ とすると (1) 式・左辺は、

$$\text{左辺} = (a - b)\,(u - v) \tag{4}$$

であって、(1) 式・右辺は、

$$\text{右辺} = a(u - v) - b(u - v) \tag{5}$$

である。

右辺の各項は (1) 式と同じ形なので

$$\text{右辺} = au - av - (bu - bv) \tag{6}$$

$$= au - av - bu + bv \tag{7}$$

である。記号は何を使っても式の意味はかわらないので u を c、v を d、に置き換え、左右に並べると、

$$(a - b)(c - d) = ac - ad - bc + bd \tag{2}$$

を得る。

(2) から (3) の導出

(2) において $c = a,\ d = b$ とする。すると左辺は

$$\text{左辺} = (a - b)(a - b) = (a - b)^2 \tag{8}$$

そして右辺は

$$\begin{aligned} \text{右辺} &= aa - ab - ba + bb \quad (9) \\ &= a^2 - 2ab + b^2 \quad (10) \end{aligned}$$

したがって (3) を得る。

6.2　総和

定義

n 個の数値があり、x_i を i 番目の値とする。全ての x_i を足し合わせた値を『総和』と呼び、$\displaystyle\sum_{i=1}^{n} x_i$ という記号であらわす。

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{11}$$

6.2.1　定数 c の総和

全ての x_i が定数 c に等しいとする。つまり $\forall x_i = c$ とする。

$$\sum_{i=1}^{n} c = \underbrace{c + c + \cdots + c}_{n} \tag{12}$$

そもそも数値は n 個あるので右辺の c も n 個ある。和を積であらわすと

$$= nc \tag{13}$$

である。

6.2.2 定数 (a) 倍された x_i の総和

x_i を定数 a 倍した値を ax_i とあらわす。ここで ax_i の総和を考える。

$$\sum_{i=1}^{n} ax_i = ax_1 + ax_2 + \cdots + ax_n \tag{14}$$

右辺において全ての項は共通因子 (a) をもつので括弧に括り掃きだすと

$$= a\left(x_1 + x_2 + \cdots + x_n\right) \tag{15}$$

括弧の中は (11) なので、

$$= a\sum_{i=1}^{n} x_i \tag{16}$$

である。

6.2.3 x_i と y_i の和の総和

二つの変数の組が n 個あるとする。変数をあらわす記号を x, y、組を識別する添え字を i とし変数の組を (x_i, y_i) とあらわす。ここで x_i と y_i の和の総和を考える。x_i と y_i の和は $x_i + y_i$ なので定義にもとづき総和をあらわすと、

$$\sum_{i=1}^{n}(x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n) \tag{17}$$

である。ここで冗長な括弧を外すと、

$$= x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n \tag{18}$$

である。

足し算は順序を変えても値は変わらないので、x_i の項と y_i の項を分けると、

$$= (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n) \tag{19}$$

である。括弧内はそれぞれ x_i，y_i の総和なので、

$$= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{20}$$

である[1)]。

1）　3 項以上の場合も同様である。

6.2.4 総和の公式

総和の定義より導かれた結果を公式として示す。

$$\sum_{i=1}^{n} c = nc \qquad ;c \text{ は定数} \tag{21}$$

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i \qquad ;a \text{ は定数} \tag{22}$$

$$\sum_{i=1}^{n} \left(x_i + y_i\right) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{23}$$

6.3　平方根とは

二乗すると a になる数を、a の平方根という。
$a>0$ のとき a の平方根は正と負の二つあり、それらの絶対値は等しい。
その正の平方根を $\sqrt{a}$ であらわす。負の平方根は $-\sqrt{a}$ である。
これら二つを合わせて $\pm\sqrt{a}$ とあらわす。
0 の平方根は 0 だけであり、$\sqrt{0}=0$ と定義する。

$$a>0 \text{ のとき}\quad \sqrt{a}^{\,2}=\left(-\sqrt{a}\right)^2=a\ ,\quad \sqrt{a}>0 \tag{24}$$

$$a=0 \text{ のとき}\quad \sqrt{0}=0 \tag{25}$$

6.3.1　b^2 の平方根

$b^2 \geqq 0$ である。$b^2 > 0$ ならば、二乗すると b^2 になる数は、$-b$ と b である。つまり、b^2 の平方根は $-b$ と b の二つある。それらのうち正の値が b^2 の正の平方根 $\sqrt{b^2}$ である。

ここで、

$$b < 0 \quad \Longrightarrow \quad -b > 0 \tag{26}$$

である。

$b^2 = 0$ ならば、$b = 0$ なので定義により $\sqrt{b^2} = b$ である。したがって、

$$b \geqq 0 \quad \Longrightarrow \quad \sqrt{b^2} = b \tag{27}$$

$$b < 0 \quad \Longrightarrow \quad \sqrt{b^2} = -b \tag{28}$$

である。これらをまとめて、

$$\sqrt{b^2} = |b| \tag{29}$$

である。

6.3.2　根号を含む式の計算 1

$a > 0,\ b > 0$ のとき

$$\left(\sqrt{a}\sqrt{b}\right)^2 = \left(\sqrt{a}\sqrt{b}\right)\left(\sqrt{a}\sqrt{b}\right) = \sqrt{a}\sqrt{b}\sqrt{a}\sqrt{b} = \left(\sqrt{a}\sqrt{a}\right)\left(\sqrt{b}\sqrt{b}\right) \tag{30}$$

$$= ab \tag{31}$$

また $\sqrt{a} > 0, \sqrt{b} > 0$ なので、$\sqrt{a}\sqrt{b} > 0$ であるから $\sqrt{a}\sqrt{b}$ は ab の正の平方根である。すなわち

$$\sqrt{a}\ \sqrt{b} = \sqrt{ab} \tag{32}$$

である。

6.3.3　根号を含む式の計算 2

$a>0,\ b>0$ のとき

$$\left(\frac{\sqrt{a}}{\sqrt{b}}\right)^2 = \left(\frac{\sqrt{a}}{\sqrt{b}}\right)\left(\frac{\sqrt{a}}{\sqrt{b}}\right) = \frac{\sqrt{a}\sqrt{a}}{\sqrt{b}\sqrt{b}} = \frac{a}{b} \tag{33}$$

また $\sqrt{a}>0, \sqrt{b}>0$ なので、$\dfrac{\sqrt{a}}{\sqrt{b}}>0$ である。したがって $\dfrac{\sqrt{a}}{\sqrt{b}}$ は $\dfrac{a}{b}$ の正の平方根である。すなわち

$$\frac{\sqrt{a}}{\sqrt{b}} = \sqrt{\frac{a}{b}} \tag{34}$$

である。

問題 II–6 – 1

$a > 0,\ b > 0$ のとき

$$\sqrt{a^2b} = a\sqrt{b}$$

が成り立つことを示せ。

解例 II–6 − 1

仮定より、$a > 0$ なので $a^2 > 0$ である。そして、同様に仮定より $b > 0$ なので、

$$\sqrt{a^2 b} = \sqrt{a^2}\sqrt{b}$$

である。$a > 0$ なので $\sqrt{a^2} = a$ である。
したがって

$$= a\sqrt{b}$$

である。

6.3.4　根号の公式

確認できた根号を含む計算の結果を公式として示す。ここで、$a > 0,\ b > 0$ とする。

$$\sqrt{a}\sqrt{b} \;=\; \sqrt{ab} \tag{32}$$

$$\frac{\sqrt{a}}{\sqrt{b}} = \sqrt{\frac{a}{b}} \tag{34}$$

$$\sqrt{a^2 b} = a\sqrt{b} \tag{35}$$

6.4　平均と分散と標準偏差

6.4.1　平均

n 個ある数値 (x_i) の総和を n で割った値を『平均』といい $\bar{x}$ であらわす。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{36}$$

両辺を n 倍すると

$$n\bar{x} = \sum_{i=1}^{n} x_i \tag{37}$$

である。

6.4.2　平均からの偏差

x_i と $\bar{x}$ の差を『平均からの偏差』という。

$$x_i - \bar{x} \tag{38}$$

『平均からの偏差』が

正の場合は x_i が平均 $\bar{x}$ より大きな値であることを意味し、

負の場合は x_i が平均 $\bar{x}$ より小さな値であることを意味する。

6.4.3　分散と標準偏差

x_i の『平均からの偏差』の二乗の平均値を分散といい V_x であらわす。

$$V_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{39}$$

分散 (V_x) の正の平方根を標準偏差と呼び、σ_x であらわす。

$$\sigma_x = \sqrt{V_x} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{40}$$

σ はギリシャ文字シグマの小文字である。

6.5　変数間の関係をあらわす指標

マーケティングでは、マーケティング変数を通じて需要へ影響を与えることができると仮定してる。つまり、需要を直接コントロールはできないが、統制可能な変数を用いて間接的にコントロールすることはできると考えている。さらに需要を予測することは困難でも、需要に密接に関連する値は予想が可能である場合も想定している。これらの状況を想定し関連する二つの値の関係を記述することを考察する。

- 単価と販売数量
- 気温とアイスクリームの販売数量

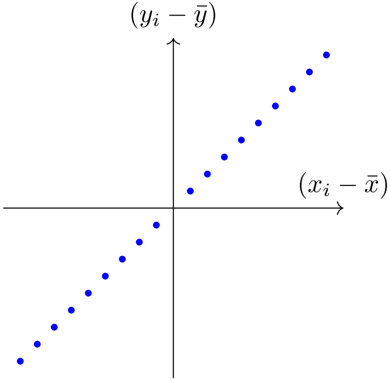
6.5.1　正の相関関係

(x_i, y_i) において

- $(x_i - \bar{x})$ と $(y_i - \bar{y})$ の符号が等しく
- 一方の絶対値が大きければ他方も大きい
- 一方の絶対値が小さければ他方も小さい

とき、『正の相関関係』があるという。

図 1　正の相関関係



6.5.2 負の相関関係

(x_i, y_i) において

- $(x_i - \bar{x})$ と $(y_i - \bar{y})$ の符号が異なり
- 一方の絶対値が大きければ他方も大きい
- 一方の絶対値が小さければ他方も小さい

とき、『負の相関関係』があるという。

図 2 負の相関関係

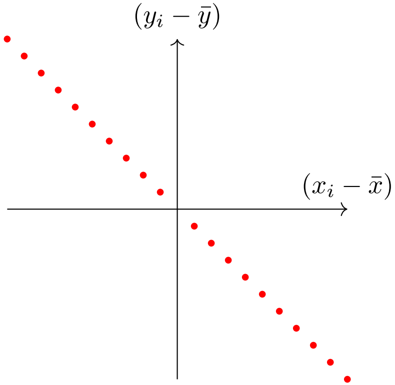


図 3 $(x_i - \bar{x})$ と $(y_i - \bar{y})$ の符号

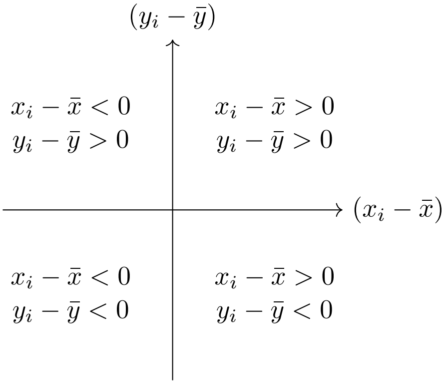
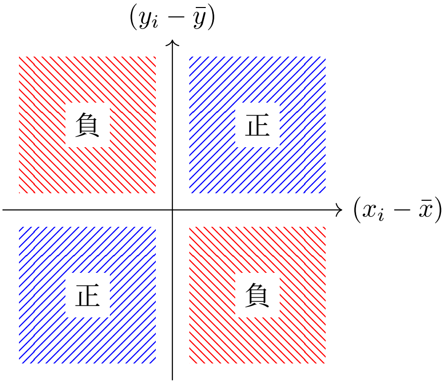


図 4 $(x_i - \bar{x})(y_i - \bar{y})$ の符号



6.5.3　共分散

定義

$(x_i - \bar{x})(y_i - \bar{y})$ の平均を『共分散』と呼び Cov_{xy} であらわす。

$$Cov_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{41}$$

共分散は二つの変数間の関係性をあらわす指標である。共分散の英語は「covariance」なので語頭の 3 文字を使って変数名としている。

数値例

表 1 はある小学校の身長, 体重,50m 走のタイムのデータである。ここで、身長を x, 体重を y,50m 走のタイムを x とすると、

$$\bar{x} = 130.783\,333\cdots \tag{42}$$

$$\bar{y} = 29.65 \tag{43}$$

$$\bar{z} = 10.066\,666\cdots \tag{44}$$

である。

表 1　身長と体重と 50m 走のタイム

i	身長	体重	タイム
1	116.5	21.3	11.6
2	122.6	24.2	10.7
3	128.5	27.7	10.0
4	133.5	31.1	9.7
5	138.6	34.4	9.4
6	145.0	39.2	9.0

図 5　身長と体重の散布図

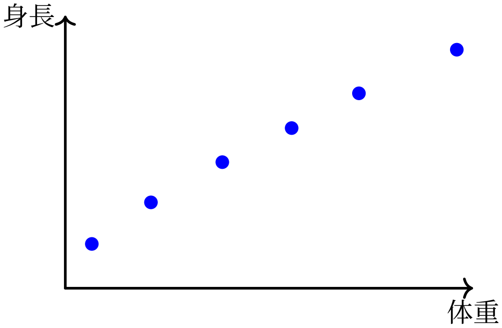
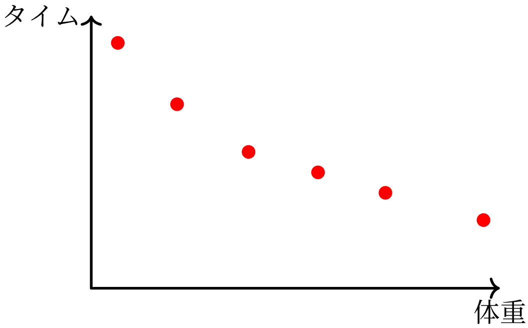


図 6　体重とタイムの散布図



x, y の平均からの偏差を求め、それぞれの積を求めたものが表 4 である。

「『平均からの偏差』の積」の平均が求めている共分散である。

$$Cov_{xy} = \frac{345.155}{6} = 57.525\,833\cdots \tag{45}$$

表 2 平均からの偏差の積の符号（正）

i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	−14.283	−8.350	119.265 833
2	−8.183	−5.450	44.599 166
3	−2.283	−1.950	4.452 500
4	2.717	1.450	3.939 166
5	7.817	4.750	37.129 166
6	14.217	9.550	135.769 166
Total			345.155 000

y, z の平均からの偏差を求め、それぞれの積を求めたものが表 3 である。

「『平均からの偏差』の積」の平均が求めている共分散である。

$$Cov_{xy} = \frac{-30.01}{6} = -5.001\,666\cdots \qquad (46)$$

表 3　平均からの偏差の積の符号（負）

i	$y_i - \bar{y}$	$z_i - \bar{z}$	$(y_i - \bar{y})(z_i - \bar{z})$
1	-8.35	1.533	$-12.803\,333$
2	-5.45	0.633	$-3.451\,666$
3	-1.95	-0.067	$0.130\,000$
4	1.45	-0.367	$-0.531\,666$
5	4.75	-0.667	$-3.166\,666$
6	9.55	-1.067	$-10.186\,666$
Total			$-30.010\,000$

問題 II–6 – 2

以下の式を展開しなさい。

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

解例 II–6 – 2

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \frac{1}{n}\sum_{i=1}^{n}(x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \frac{1}{n}\sum_{i=1}^{n}x_i\bar{y} - \frac{1}{n}\sum_{i=1}^{n}\bar{x}y_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{y}\frac{1}{n}\sum_{i=1}^{n}x_i - \bar{x}\frac{1}{n}\sum_{i=1}^{n}y_i + \frac{1}{\cancel{n}}\cancel{n}\bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{y}\bar{x} - \cancel{\bar{x}\bar{y}} + \cancel{\bar{x}\bar{y}} \\
&= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\bar{y}
\end{aligned}
$$

数値例

$$Cov_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{47}$$
$$= \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\bar{y} \tag{48}$$

である。この結果を利用して表 1 をもとに x, y 間の共分散を求めると、

$$Cov_{xy} = 3935.25 - 130.78 \times 29.65 \tag{49}$$
$$= 57.525\,833\cdots \tag{50}$$

表 4 $\boldsymbol{x_i y_i}$

i	x_i	y_i	$x_i y_i$
1	116.5	21.3	2,481.45
2	122.6	24.2	2,966.92
3	128.5	27.7	3,559.45
4	133.5	31.1	4,151.85
5	138.6	34.4	4,767.84
6	145.0	39.2	5,684.00
平均	130.78	29.65	3,935.25

同様に y, z 間の共分散を求めると、

$$Cov_{yz} = 293.48 - 29.65 \times 10.07 \tag{51}$$

$$= -5.001\,666\cdots \tag{52}$$

表 5 $y_i z_i$

i	y_i	z_i	$y_i z_i$
1	21.30	11.60	247.08
2	24.20	10.70	258.94
3	27.70	10.00	277.00
4	31.10	9.70	301.67
5	34.40	9.40	323.36
6	39.20	9.00	352.80
平均	29.65	10.07	293.48

6.5.4 相関係数

定義

共分散を x_i,y_i それぞれの標準偏差 σ_x, σ_y で割った値を『相関係数』と呼び ρ_{xy} とあらわす。ここで、$\sigma_x \neq 0$, $\sigma_y \neq 0$ とする。

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \ \sigma_y} \tag{53}$$

ρ はギリシャ文字ローの小文字である。

問題 II–6 – 3

表の値を用いて、ρ_{xy}，ρ_{xz}，ρ_{yz} を求めなさい。

値の算出に際して何らかのアプリを使用することは自由である。

i	x_i	y_i	z_i
1	116.5	21.3	11.6
2	122.6	24.2	10.7
3	128.5	27.7	10.0
4	133.5	31.1	9.7
5	138.6	34.4	9.4
6	145.0	39.2	9.0

数値例

表 1 をもとに、身長, 体重,50m 走のタイムの共分散と相関係数を求めてみる。身長を x, 体重を y, 50m 走のタイムを z とすると、それぞれの変数間の共分散は

$$Cov_{xy} = 57.525\,833\cdots \tag{54}$$

$$Cov_{xz} = -8.050\,555\cdots \tag{55}$$

$$Cov_{yz} = -5.001\,666\cdots \tag{56}$$

である。

また、相関係数は

$$\rho_{xy} = 0.997\,213\cdots \tag{57}$$

$$\rho_{xz} = -0.976\,679\cdots \tag{58}$$

$$\rho_{yz} = -0.958\,592\cdots \tag{59}$$

である。

正の相関が強い場合、相関係数は 1 に近い値を示し、負の相関が強い場合、–1 に近い値を示す。

問題 II–6 – 4

$$Cov_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

とする。

ここで、

$$n > 1\ ,\quad \sigma_x > 0\ ,\quad \sigma_y > 0$$

とする。

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x\ \sigma_y}$$

を計算しなさい。ただし解答に際して、『平均からの偏差』は展開せず、また分母を有理化する必要はない。

解例 II–6 – 4

$$
\begin{aligned}
\frac{Cov_{xy}}{\sigma_x\sigma_y} &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \\
&= \frac{\cancel{\frac{1}{n}}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\cancel{\frac{1}{n}}\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}
\end{aligned}
$$

6.5.5　相関係数の範囲

相関係数 ρ_{xy} はその範囲が

$$-1 \leqq \rho_{xy} \leqq +1 \tag{60}$$

の範囲に限定される。この性質により 2 変数間の関係をあらわす指標として極めて有用である。

6.6 まとめ

- 根号を含む演算を確認した。
- 二つの変数 (x_i, y_i) 間の関わり合いの指標は『共分散 (Cov_{xy}) 』と『相関係数 (ρ_{xy}) 』がある。
- 「『平均からの偏差』の積」の平均が共分散 (Cov_{xy}) である。
- 共分散 Cov_{xy} をそれぞれの標準偏差 $(\sigma_x,\ \sigma_y)$ で割った値が相関係数 ρ_{xy} である。
- 相関係数 ρ_{xy} は -1 以上 $+1$ 以下の範囲に収まるので有用である。