第一章　データの性質

劉慶豊[1]

小樽商科大学

November 18, 2009

[1]E-mail:qliu@res.otaru-uc.ac.jp, URL:http://www.otaru-uc.ac.jp/~qliu/

データの性質はデータそのものを眺めるだけでは良く分からない。データの特性を分かりやすく示すためには、さまざまな代表値が利用される。

Income

平均と分散

標本平均

- **変数xに関するデータ$\{x_1, x_2, \cdots, x_n\}$が与えられているとする．変数$x$の標本平均は**

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

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{2}$$

$$x_1 = 3, x_2 = 7, x_3 = 8, x_4 = 12, x_5 = 20.$$

$$\frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}(3 + 7 + 8 + 12 + 20) = 10$$

加重平均

- **標本の重要さに応じて異なった重みを付けて求めた平均。**

$$\sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n, \sum_{i=1}^{n} w_i = 1 \tag{3}$$

重み（ウエイト）w_iが非負で総和が1である。

- $$\sum_{i=1}^{5} w_i x_i = \frac{1}{9}3 + \frac{2}{9}7 + \frac{3}{9}8 + \frac{2}{9}12 + \frac{1}{9}20 = \frac{85}{9} = 9.44$$

- **例、学生の全科目の平均成績の計算を行うとき、科目の重要性に応じて異なったウエイトを付ける場合がある。ウエイトとして、国語1/4英語1/4数学1/4生活1/8図画工作1/8とする。ある学生の得点が国語90英語80数学95生活85点、図画工作100点、**
- **この場合の加重平均による平均成績は**

$$\sum_{i=1}^{5} w_i x_i = \frac{1}{4} \times 90 + \frac{1}{4} \times 80 + \frac{1}{4} \times 95 + \frac{1}{8} \times 85 + \frac{1}{8} \times 100 \approx 89$$

移動平均

- **時系列データ** $\{x_1, x_2, \cdots, x_t\}$ **によく使われる。** t **時点の近くの値で** t **時点の平均を計算する。**
- 3**項平均：**

$$\bar{x}_t = \frac{x_{t-1} + x_t + x_{t+1}}{3}$$

- 4**項平均**

$$\bar{x}_t = \frac{x_{t-1} + x_t + x_{t+1} + x_{t+2}}{4}$$

- **データ全体に関して移動平均を取ると、新しいデータの系列が出来る：３項平均の場合** $\{\bar{x}_2, \bar{x}_3, \cdots, \bar{x}_{n-1}\}$**、４項平均の場合** $\{\bar{x}_3, \bar{x}_4, \cdots, \bar{x}_{n-2}\}$**。**

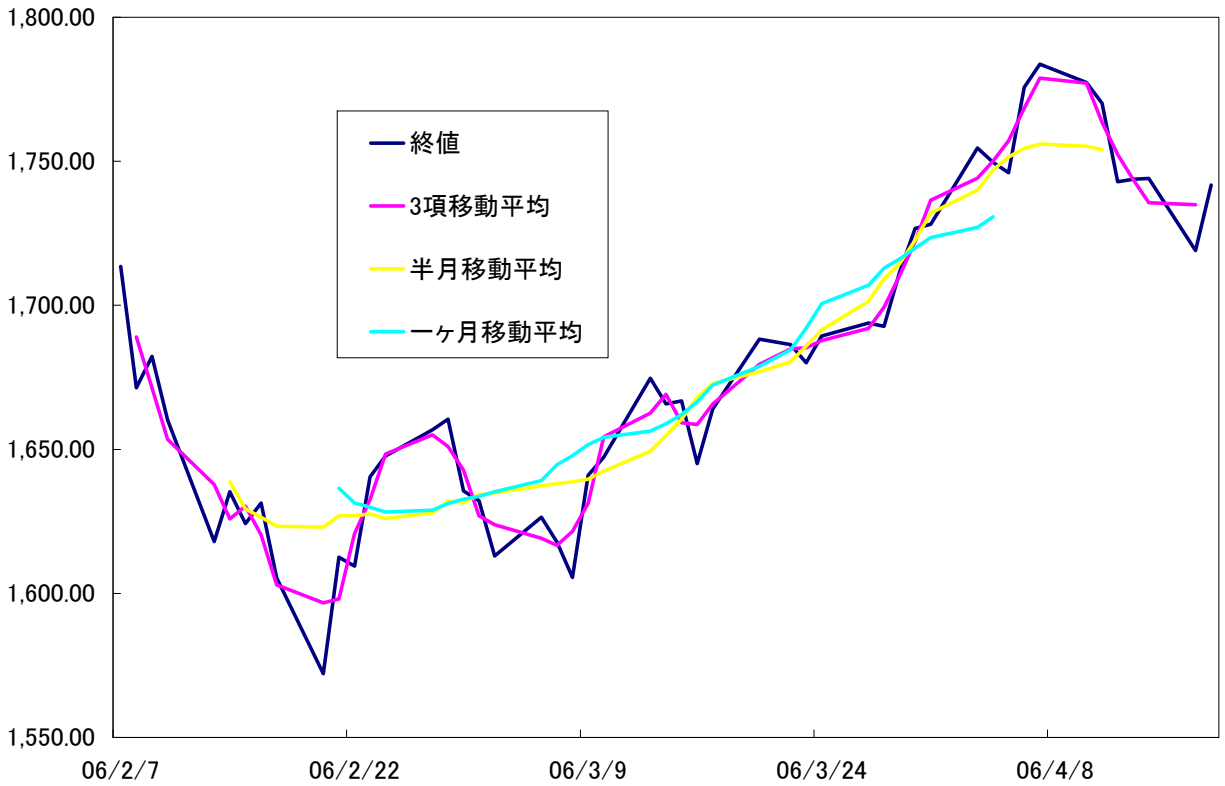
移動平均の例

- 株や為替レートチャートなどによく使われる。移動平均を項数が多ければ滑らかになり（ギザギザが消えていく）、より長期的な動きを反映する。

日付	終値	3項移動平均	5項移動平均	7項移動平均
18/04/2006	1741.75			
17/04/2006	1719.05	1734.96		
14/04/2006	1744.07	1735.63	1738.31	
13/04/2006	1743.77	1743.58	1743.99	1748.44
12/04/2006	1742.89	1752.28	1755.65	1754.43
11/04/2006	1770.18	1763.47	1763.58	1762.52
10/04/2006	1777.34	1777.08	1769.96	1762.80
07/04/2006	1783.72	1778.91	1770.59	1763.64
06/04/2006	1775.67	1768.48	1766.49	1765.32
05/04/2006	1746.05	1757.12	1761.95	1759.32
04/04/2006	1749.65	1750.11	1750.83	1752.08
03/04/2006	1754.64	1744.15	1741.04	1741.77
31/03/2006	1728.16	1736.49	1734.13	
30/03/2006	1726.68	1722.13		
29/03/2006	1711.54			

TOPIXの日次データの移動平均

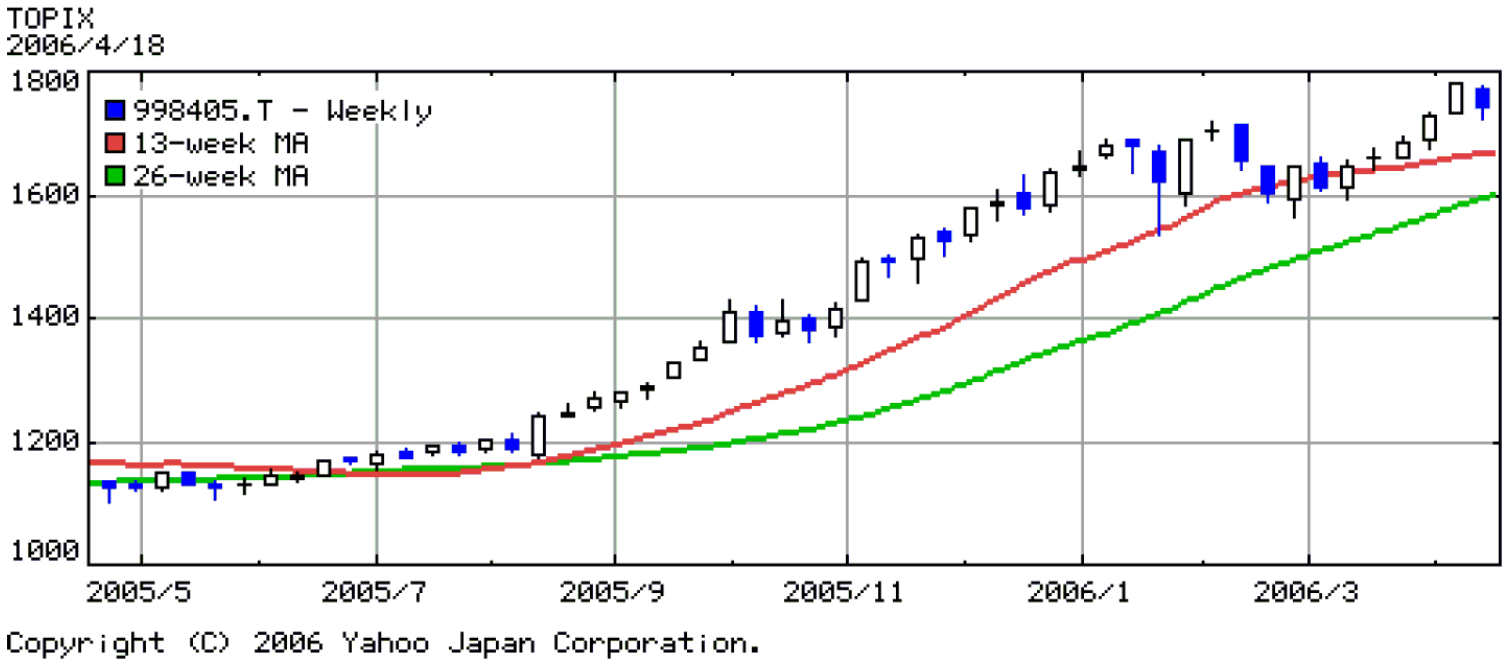
移動平均の例（続き）



TOPIXの日次データの移動平均

移動平均の例（続き）

- 移動平均の動きを見て株価の予測を行っている投資家がいる。



中央値（メディアンmedian）

- データを小さいものから順に並べて、その真ん中にある値は中央値である。
- 例

$$X = \{4, 7, 2, 30, 9, 7, 1\}$$

$$X^* = \{1, 2, 4, 7, 7, 9, 30\}$$

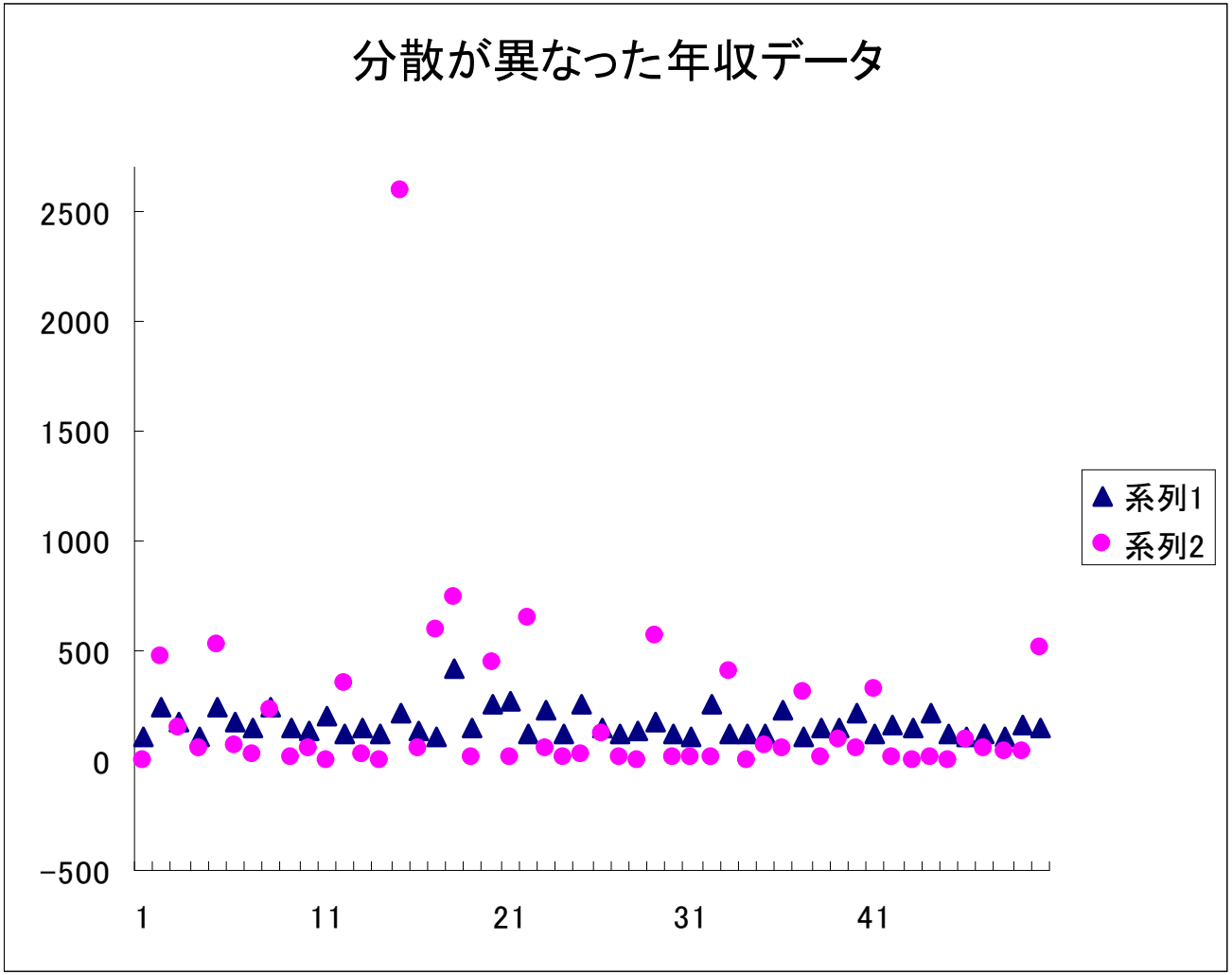
7が真ん中に位置するので中央値である。

- 中央値は算術平均より異常値の影響を受けにくい。

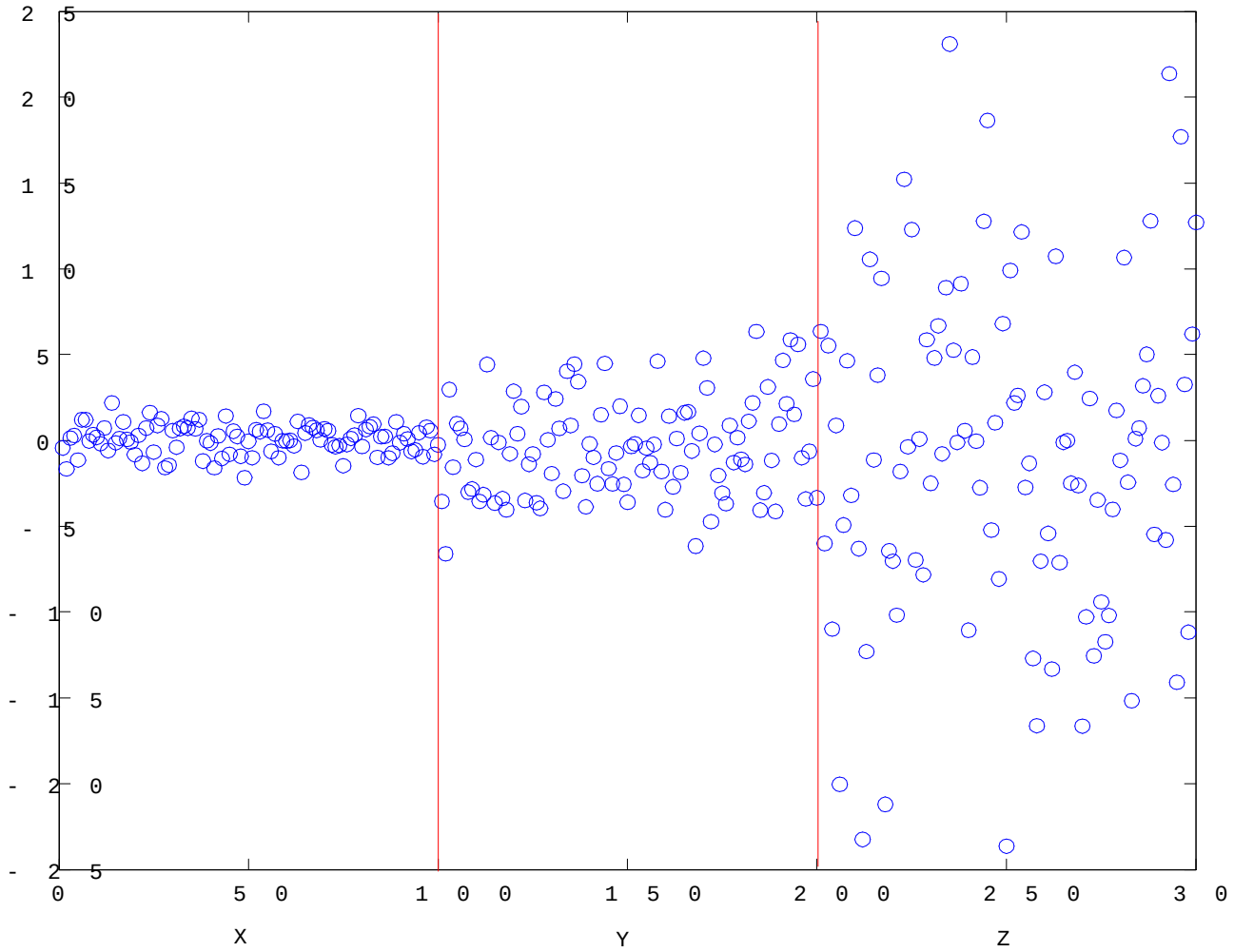
中央値の応用例

- 例、某社の株価の先週一週間月曜日から金曜日までの五日間の株価はそれぞれ$100, 90, 10,\ \ 110, 110$円となっているとする。水曜日の10円の高値はその会社の収益実績に関する噂によるもので、異常値である。平均を計算すれば、84になるが、中央値は100となっている。明らかに、平均値は異常値の影響を大きく受けていて、株価の平均でこの会社の実力を評価すると過小評価につながる。一方では中央値の方はより忠実にこの会社の実力を反映している。

標本分散


分散が異なった年収データ
2500
2000
1500
1000
500
0
−500
1
11
21
31
41
系列1
系列2

標本分散



標本分散

- **分散はデータが平均からの乖離の具合を図る尺度で散らばりの具合を表す指標である。**

$$S_x^2 = S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{4}$$

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = (x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2 \tag{5}$$

標本分散の分解公式

- $$S_x^2 = S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

Proof.

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x})^2 &= (x_1^2 - 2x_1\overline{x} + \overline{x}^2) + (x_2^2 - 2x_2\overline{x} + \overline{x}^2) + \cdots \\ &\quad + (x_n^2 - 2x_n\overline{x} + \overline{x}^2) \\ &= (x_1^2 + x_2^2 + \cdots + x_n^2) - 2(x_1 + x_2 + \cdots + x_n)\overline{x} \\ &\quad + (\overline{x}^2 + \overline{x}^2 + \cdots + \overline{x}^2) \\ &= (x_1^2 + x_2^2 + \cdots + x_n^2) - 2n\overline{x}^2 + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2\end{aligned}$$

□

例題

- **データ** $x = \{3, 7, 8, 12, 20\}$

$$\bar{x} = 10$$

- $$\frac{1}{4}(7^2 + 3^2 + 2^2 + 2^2 + 10^2) = 41.5$$

- $$\frac{1}{4}\{(3^2 + 7^2 + 8^2 + 12^2 + 20^2) - 5 \times 10^2\} = 41.5$$

標本標準偏差

- 標本標準偏差は標本分散の正の平方根で、確率変数x標本標準偏差はS_xで表される

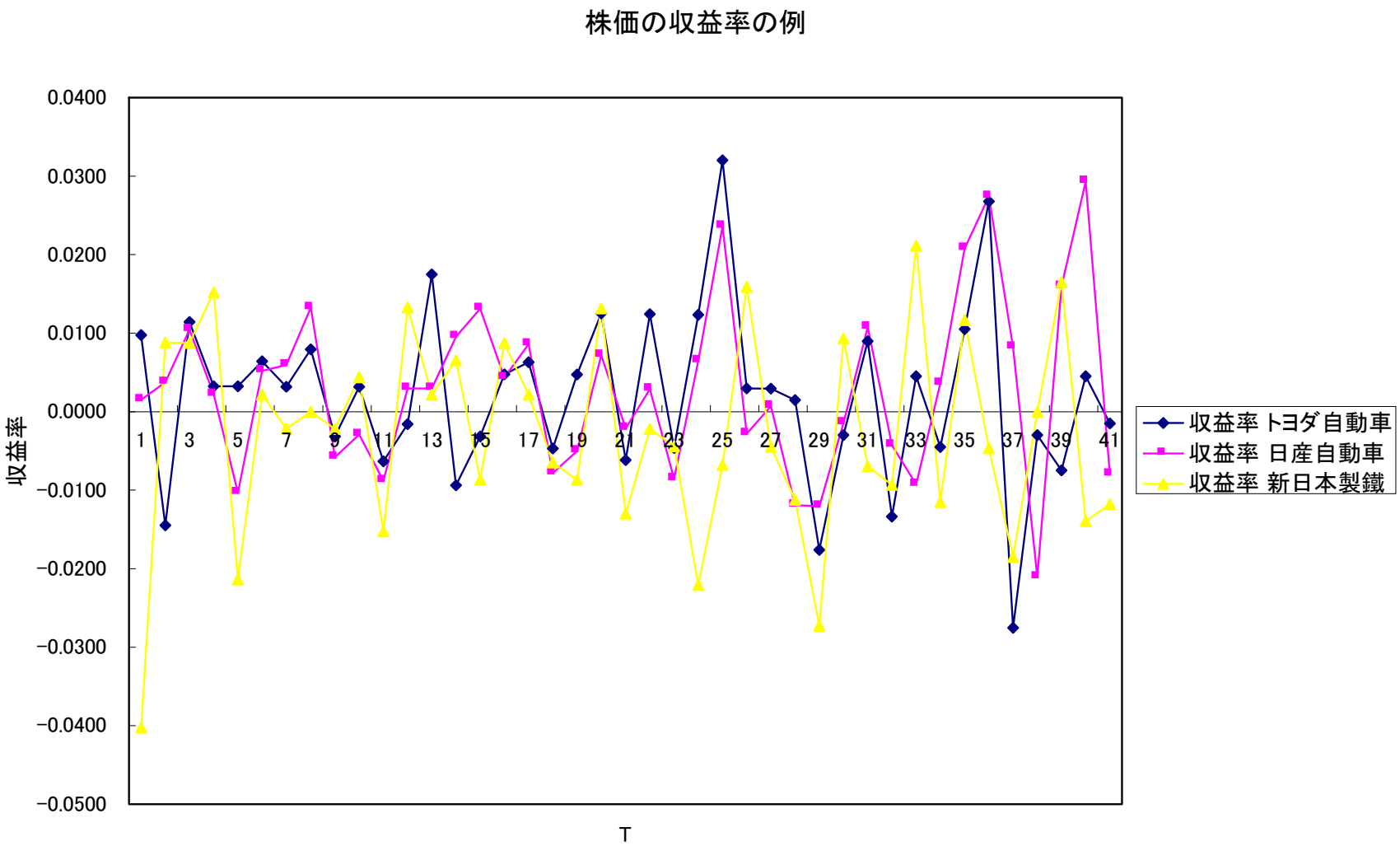
$$S_x = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- 標本標準偏差は分散と同様にデータの散らばりの具合を表す。

分散や標準偏差の応用例

- トヨダ自動車、日産自動車、新日本製鐵の株式の収益率の平均と分散はそれぞれ、(0.002, 0.0001), (0.003, 0.0001) と (-0.003, 0.0002) である。
- 期待値が高く、分散（リスク）が小さいものが良いと言う観点から、この中一番パフォマンスの良いのがトヨダ自動車である。株価の収益率の標準偏差がボラティリティと呼ばれ、リスク評価の尺度となっている。

分散や標準偏差の応用例（続き）



チェビシェフの不等式

Theorem (チェビシェフの不等式)

全観測値の中 k シグマ区間と呼ばれる以下の区間

$$(\bar{x} - k \times \text{標本標準偏差}, \quad \bar{x} + k \times \text{標本標準偏差})$$

に含まれない観測値の割合は $(1/k^2)$ 以下である。含まれる観測値の割合は $(1 - 1/k^2)$ 以上である

k シグマ区間という呼び方は慣例としてギリシャ文字 σ で母集団の標準偏差を表すためである。

チェビシェフの不等式（続き）

- 1シグマ区間（平均ー標本標準偏差，平均＋標本標準偏差）は少なくともー つ以上の観測値を含む
- データ$\{3, 7, 8, 12, 20\}$の例、平均10、標準偏差は6.4, 1シグマ区間は$(10 - 6.4, 10 + 6.4) = (3.6, 16.4)$
- 2シグマ区間は3/4以上の観測値を含む

（平均ー$2 \times$標本標準偏差，平均＋$2 \times$標本標準偏差）

- 3シグマ区間は殆どの観測値を含む

（平均ー$3 \times$標本標準偏差，平均＋$3 \times$標本標準偏差）

表1-1

テキスト表1-1

例1.1

- **所得の平均は** $210.6/18 = 11.7$**、分散は** $1/17 \times (4251.8 - 1/18 \times 210.6^2) = 105.16(1000\$^2)$,
- **寿命の平均は** $1270/18 = 70.6$**、分散は** $1/17 \times (90430 - 1/18 \times 1270^2) = 48.5$**歳**^{2},
- **所得の**1**シグマ区間** $(11700 - 10250, 11700 + 10250) = ($**−1450 ドル**, **21950 ドル**$)$
- **所得の**2**シグマ区間** $(11700 - 2 \times 10250, 11700 + 2 \times 10250) = ($**−8800 ドル**, **32200 ドル**$)$
- **寿命の**1**シグマ区間** $(70.6 - 7, 70.6 + 7) = (63.6, 77.6)$
- **寿命の**2**シグマ区間** $(70.6 - 2 \times 7, 70.6 + 2 \times 7) = (56.6, 84.6)$

データの標準化

- **平均を引いて、標準偏差で割る**

$$z_1 = \frac{x_1 - \overline{x}}{s_x}, \quad \cdots, \quad z_n = \frac{x_n - \overline{x}}{s_x} \tag{6}$$

- **標準化されたデータの平均は0、分散が1となる**

Proof.

$$\frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \overline{x}}{s_x} \tag{7}$$

$$= \frac{1}{s_x} \times \frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x} \tag{8}$$

$$= 0 \tag{9}$$

□

続き

Proof.

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n} z_i^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)^2 \\ &= \frac{1}{s_x^2}\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\overline{x})^2 \\ &= 1\end{aligned} \tag{10}$$

$$\overline{x} - ks_x < x_i \leq \overline{x} + ks_x$$

$$-k < z_i \leq k$$

□

表1-2

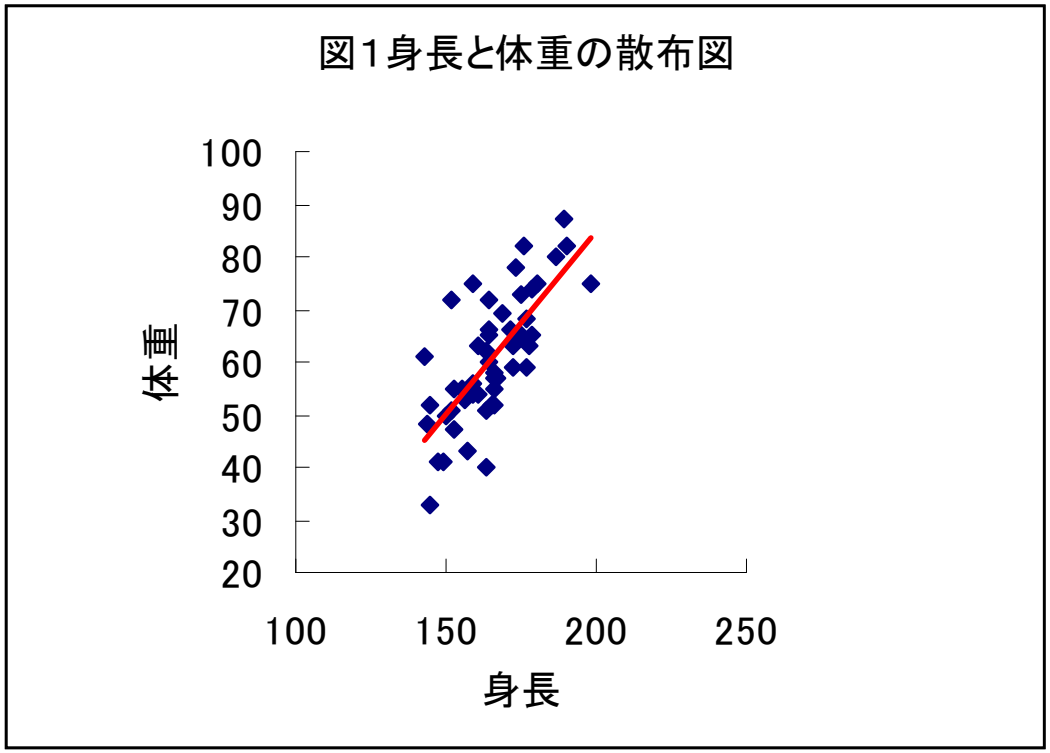
テキスト表1-2
テキスト図1-2

二変数の間の関係、散布図、分割表と相関係数

- 平均や分散などの統計量は一変数の特性を明らかにする。
- 二変数の場合、当然一変数に使われた手法や統計量は二つの変数におのおの適用できるが、二つの変数の間の関係を明確にするために、一変数の場合と違った手法や統計量が必要となる。

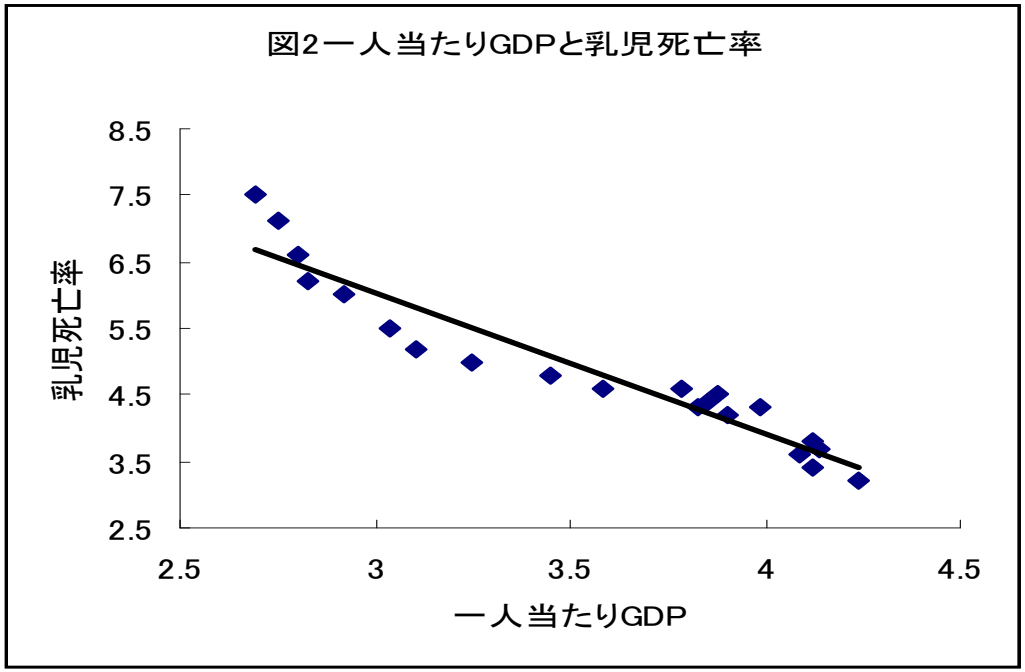
散布図

- 二つの変数をそれぞれ縦軸と横軸にして使ったグラフは散布図となる。
- 散布図から二つの変数間の線形関係が見えてくる。身長が高ければ体重が重いはずである。散布図で見るとこのような身長と体重の関係は右上がりのラインを一本引けるように見える。



散布図続き

- 日本の1980年から2000年までの一人当たりGDP(国内総生産）と乳児死亡率のデータの散布図。一人当たりGDPの成長と伴に、乳児死亡率も下がっていることが分かる。右下がりの直線を引けるように見える。



データの出所：GDPのデータは内閣府SNA、乳児死亡率は厚生労働省大臣官房統計情報部人口動態・保健統計課、人口のデータは総務

標本共分散と標本相関係数

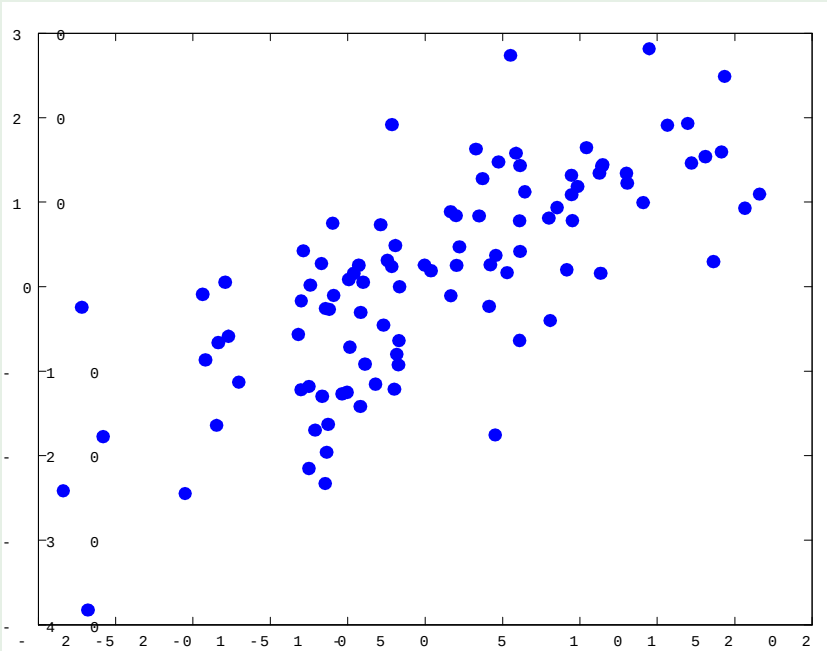
- 共分散と相関係数は二変数間の相関関係を示す統計量である。二つの変数XとYに関して、S_{xy}でXとYの共分散をρ_{xy}でXとYの相関係数を表す。
- 共分散

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

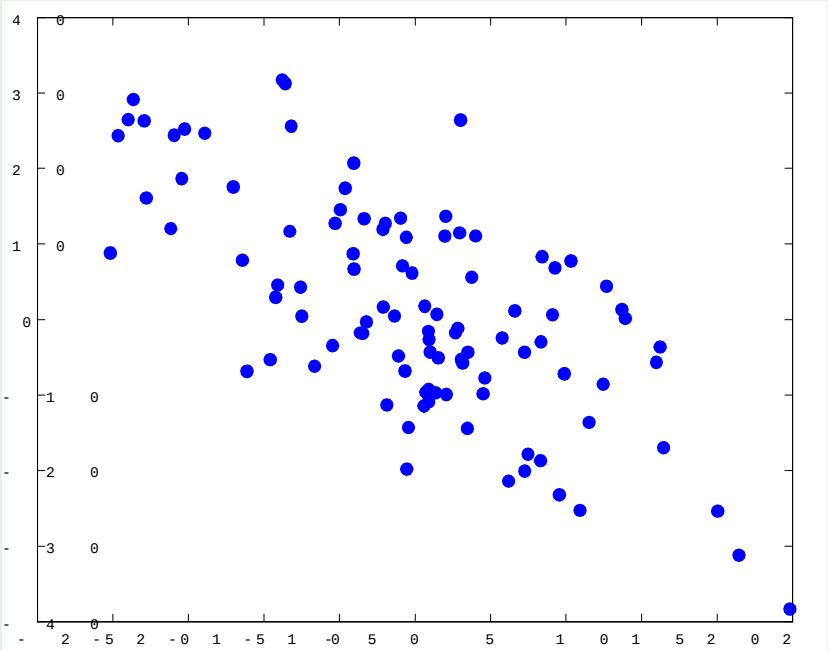
標本共分散（続き）

Example

図３の中に共分散がプラスとマイナスの例を示している。



正の共分散を持つ



負の共分散を持つ

図3 異なった共分散を持ったデータの散布図

標本相関係数

Definition

相関係数は標準化された共分散である。

$$\begin{aligned}\rho_{xy} &= \frac{S_{xy}}{S_x S_y} \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}\end{aligned}$$

- **一つ重要な性質：任意の二つの変数の相関係数は**

$$-1 \leq \rho \leq 1$$

である。

異なった相関係数を持ったデータの散布図

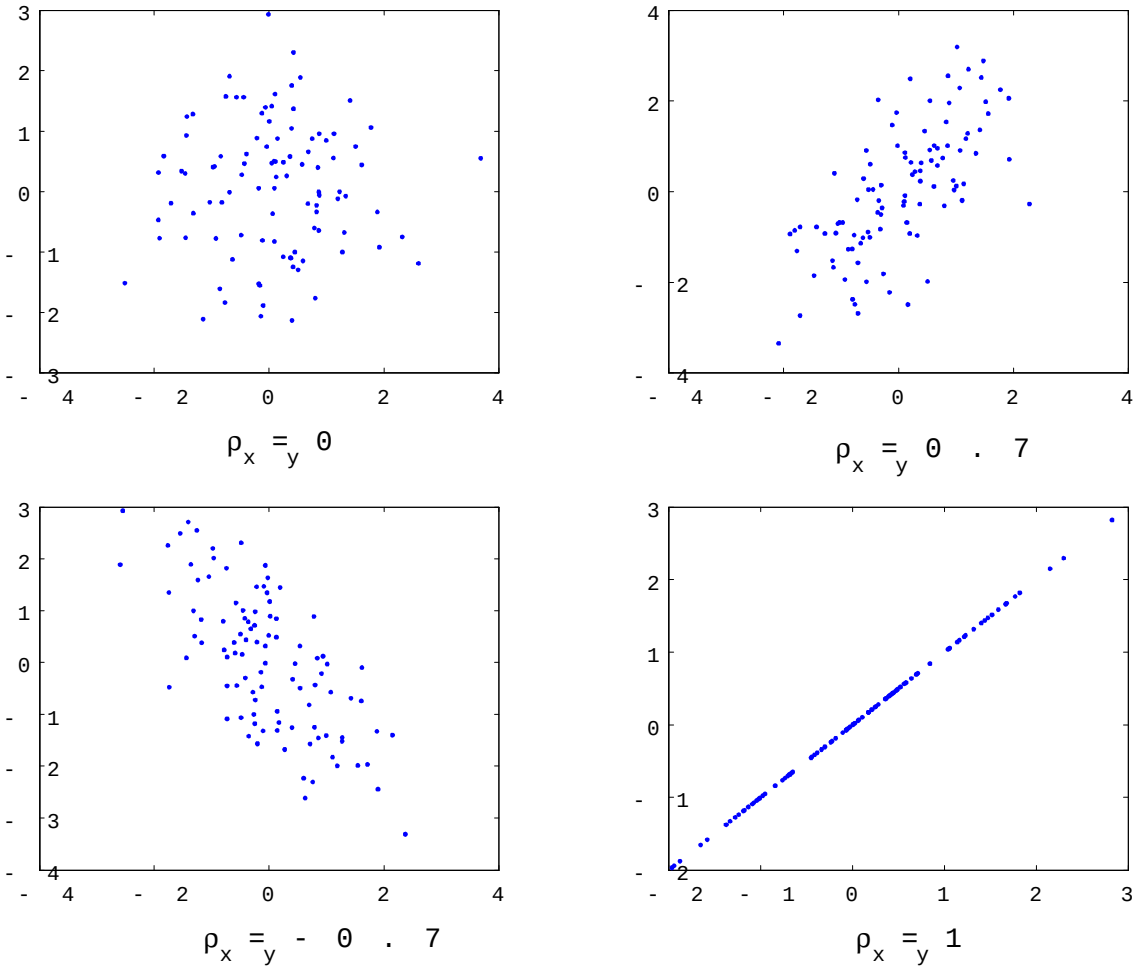


図４異なった相関係数を持ったデータの散布図

例

Example

肥料と農産物の出来高の関係を相関係数で見ることが出来る。(以下のデータは架空のデータ)。

田んぼのID	a	b	c	d	e	f	g	h	i	j
肥料の量(g)	10	15	20	25	30	35	40	45	50	55
出来高(kg)	6	15	19	18	7	19	24	22	34	35

続き

Example

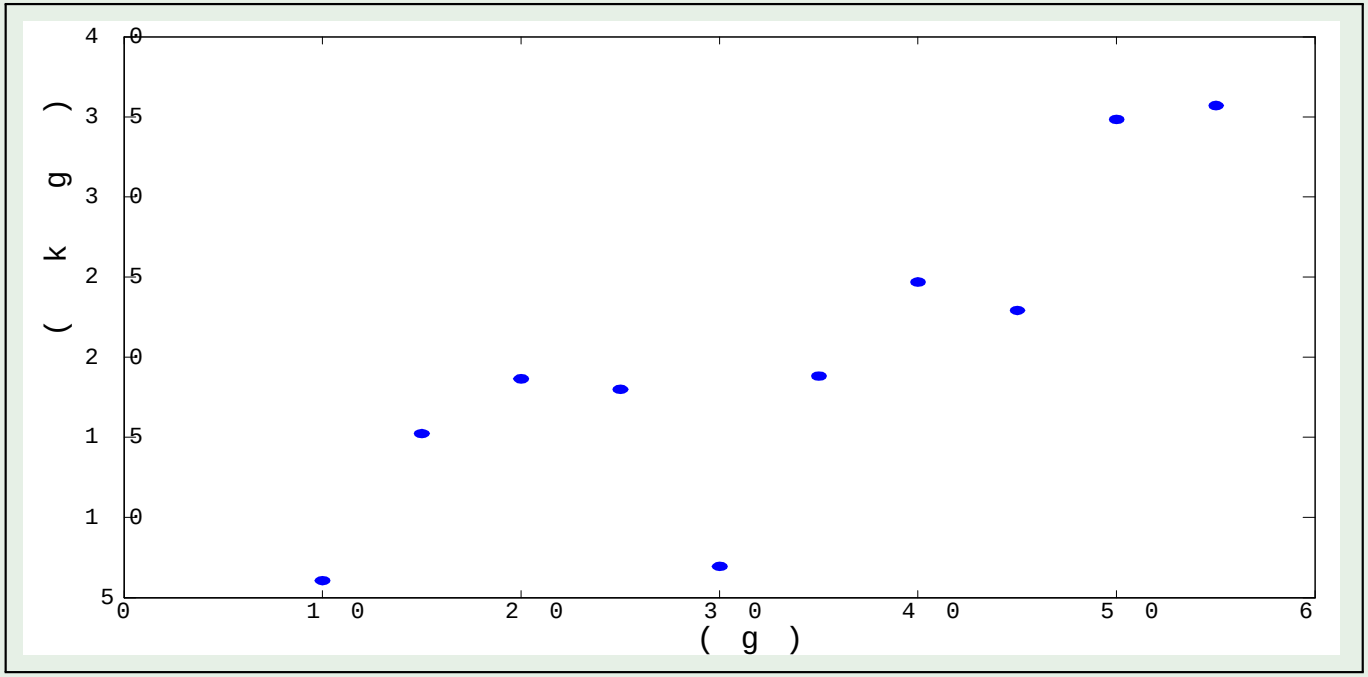


図5 肥料と農産物の出来高との関係

相関係数を計算したら、$\rho_{xy} = 0.84$従って、この種の肥料は農産物の増産に大きく寄与していると結論付けられる。