第一章　記述統計

劉慶豊[1]

小樽商科大学

May 11, 2012

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

データの種類

1. 質的なデータ：性質を表すデータ、数値だけでは表せない。
   例：赤か緑か、男か女か、大学生か高校生か・・・
2. 量的なデータ：数量を表すデータ、数値で表せる。
   例：降水量は何ミリ、体重は何キロ、売り上げは何円・・・

母集団と標本

1. 母集団：分析したい対象の全体を母集団と呼ばれる。
   例：このクラス全員の身長、日本国民全員の個人所得$\cdots$
2. 標本：母集団の一部。
   例：このクラスの中の10人の身長、抽選で選んだ200名日本人の個人所得$\cdots$
3. 全標本（センサスcensus）母集団の全体が観測されたとき、観測値の集合を全標本という。
4. 観測個数（標本サイズ）観測値の数。
   例：所得を変数xで示す。一部の国民の所得の計測値を$x_1, x_2, \cdots, x_n$と記す。計測値の集まり$S = \{x_1, x_2, \cdots, x_n\}$が標本。調べた国民の人数$n$が観測個数（標本サイズ）。

データを収集する方法

1. 全数調査：母集団全体に関して調査する。
2. 標本調査：選んだ標本に関して調査する。

度数分布表とヒストグラム

- 大学生男子50人の身長のデータの度数分布表

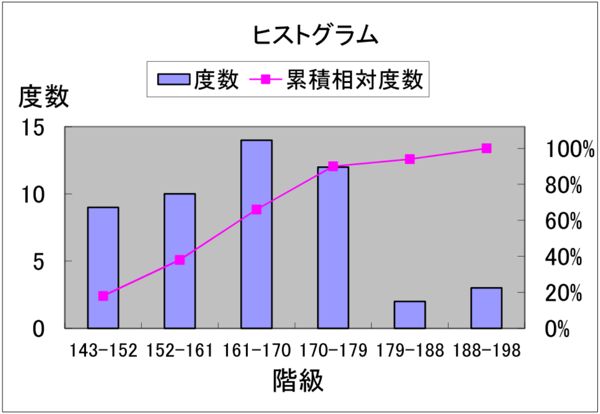
分布表の例

階級	度数	累積度数	相対度数	累積相対度数
143–152	9	9	18%	18%
152–161	10	19	20%	38%
161–170	14	33	28%	66%
170–179	12	45	24%	90%
179–188	2	47	4%	94%
188–198	3	50	6%	100%

度数 各階級に入っているデータの数.
相対度数 度数/全体のデータ数。
累積度数 下の階級からの度数の合計。
相対累積度数 累積度数/全体のデータ数。

ヒストグラム

- 度数分布表を棒グラフにして、視覚的にもっと分かりやすくなる。このような各階級の度数を棒グラフにしたものをヒストグラムという。紫の線は累積相対度数その値は右の縦軸に対応する。



代表値

代表値 代表値はデータを根拠に計算したデータの特性を代表できる数値である。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。

- 調査などで収集してきたデータの特性を代表値でまとめる。全数調査の場合では、代表値はそのまま母集団（分析対象のすべて）の特性を表す。標本（母集団の一部分）調査の場合、標本の代表値を計算して、母集団の特性を推測するために利用する。

記述統計 データの特性を代表値、度数分布表、ヒストグラムなどで表す、記述する統計学である。

推測統計 様々な数学のツールを利用して標本のデータを分析し、母集団の性質を推測する統計学である。

統計的推測 統計学では標本を使って母集団全体を理解しようとする。このための方法を統計的推測という。図1.1。

- 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（算術平均）は

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

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

- 標本平均の性質

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

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

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

切落し平均（刈り込み平均）

- データの中値の大きいものや小さいものを切り落としてから平均を計算する。異常に大きいまたは小さい値が全体に与える必要以上の影響を取り除く役割がある。異常値に対して頑健である。
- 例：国際試合の不正を防ぐため、それぞれ違う国から来た5人の裁判がいる場合、点数の最大値と最小値を切り落として合計（$\bar{x} \times 3$）を得点とする。

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。
$$\sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n, \sum_{i=1}^{n} w_i = 1 \qquad (3)$$
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$\{\bar{x}_2, \bar{x}_4, \cdots, \bar{x}_{n-2}\}$。

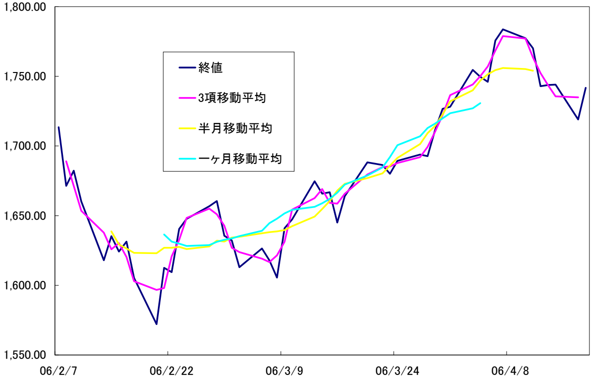
移動平均の例

- 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

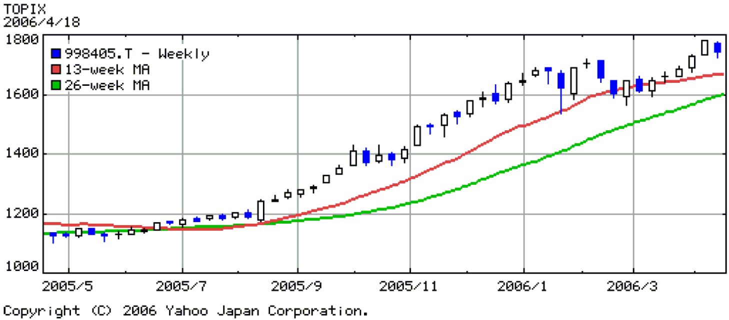
移動平均の例（続き）



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。



幾何平均

$$\sqrt[n]{\text{最終増加倍率}}$$

- 幾何平均はよく平均増加倍率の計算に使われる。例：個人所得3年間の間で各年それぞれ$r_1 = 23\%, r_2 = 27\%, r_3 = 28\%$で増加したとする、3年間を通してもとの約2倍に増加した

$$(1 + r_1)(1 + r_2)(1 + r_3) \approx 2$$

- 平均増加倍率は

$$\begin{aligned} R &= \sqrt[3]{(1 + r_1)(1 + r_2)(1 + r_3)} \\ &= \sqrt[3]{2} \approx 1.26 \end{aligned}$$

  となる。
- すなわち、平均的に毎年前の年の1.26倍まで増加すれば、3年で役2倍になる。確かめると$1 \times 1.26 \times 1.26 \times 1.26 \approx 2$.
- 例題：テキストP26。

中央値（メディアンmedian中位数）

- 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。
- データ数が偶数となっている場合は間中にある二つの観測値の平均を中央値とする。
- 中央値は安定した尺度、算術平均より異常値の影響を受けにくい。
- 異常値は他の観測値と比べて異常に大きいや異常に小さい値。

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円の安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。

四分位点

- **データを小さいものから順に並べて、**25% **目の値を第１四分位点、**75% **目の値を第**3**四分位点と呼ぶ。第**2**四分位点は**50% **目の点なので中央値となる。**
- **例：データは**$\{3, 5, 8, |10, 11, 12, |21, 22, 24, |56, 90, 92\}$, **第１四分位点**$= (8+10)/2 = 9$, **第**3**四分位点**$= (24+56)/2 = 40$ **となる、中央値**$= (12+21)/2 = 15.5$。

最頻値（モードmode）

最頻値 度数が最大な値が最頻値である。同じことで、出現頻度が一番高い値である。

Example

$$X = \{2, 3, 2, 4, 6, 4, 6, 6, 7\}$$

とする、6の度数（出現頻度、出現した回数）が一番高く（高く、多く）、3になっているので、6が最頻値である。

範囲（レンジ range）

レンジ データの最大値と最小値との差。同じくデータの取る値の範囲で理解してもいい。

$$R = \max(X) - \min(X)$$

Example

$$X = \{2, 3, 2, 4, 6, 4, 6, 6, 7\}$$

Xの最小値は2で、最大値は7なので、レンジ$= 7 - 2 = 5$。

- レンジはデータ散らばりの尺度の一つである。データの範囲をあらわしているが、レンジの範囲内でデータはどういう具合で散在しているのかに関する情報を含んでいない。

平均偏差（mean deviation）

平均偏差 $X = \{x_1, x_2, \cdots, x_n\}$ とし、平均偏差は

$$d = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n}$$

ただし $\bar{x}$ は算術平均である。

- 数式での定義しか出来ないが、解釈すれば、平均偏差は個々のデータが平均からの乖離を全データに渡って評価する一つの統計量である。

Example

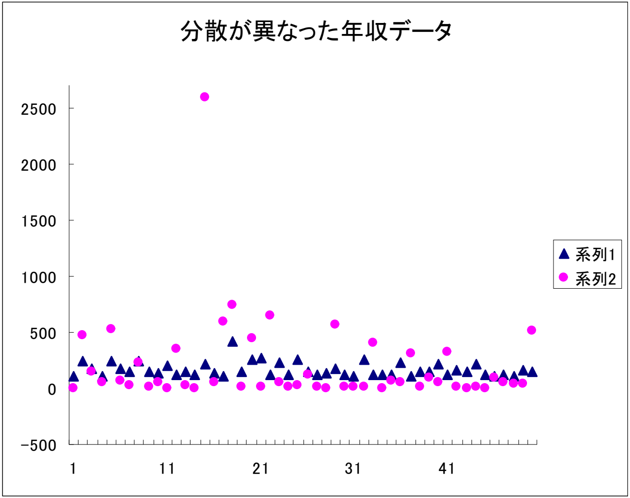
$X = \{x_1, x_2, \cdots, x_{10}\} = \{2, 3, 2, 4, 6, 4, 6, 6, 3, 4\}$ **とする、**

$$\bar{x} = \frac{2+3+2+4+6+4+6+6+7}{10} = 4$$

$$\begin{aligned} d &= \frac{|2-4| + |3-4| + |2-4| + \cdots + |7-4|}{10} \\ &= \frac{2+1+2+0+2+0+2+2+3}{10} = \frac{7}{5} \end{aligned}$$

意味としては、各データは平均的に $\bar{x}$ **から** $7/5$ **離れている。**

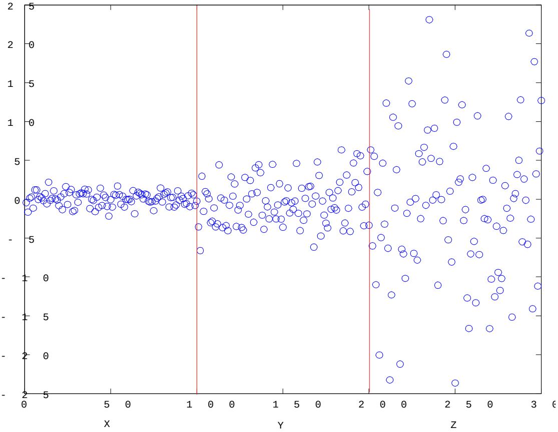
標本分散



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

- $X = \{x_1, x_2, \cdots, x_{100}\}, Y = \{y_1, y_2, \cdots, y_{100}\}, Z\{y_1, y_2, \cdots, y_{100}\}$ の平均は0でそれぞれの分散1, 3, 10である。グラフで見ると三者の違いは明らかである。分散が大きいほどデータの広がり（散らばり）の程度が大きくなる。



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。
- 以下の$1/n$を掛ける定義があるが

$$S_x^2 = S_{xx} \tag{4}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{5}$$

推測統計のため以下の$1/(n-1)$を掛ける定義の性質がいいのでこれから以下の定義を利用する

$$S_x^2 = S_{xx} \tag{6}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{7}$$

説明はテキスト146ページを参照。

標本分散の分解公式

- **標本分散の分解**

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

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

- **証明**

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

標本分散の分解公式

- 記述統計の場合よく $n-1$ の代わりに n を使って分散を以下の公式で計算する

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

そのとき、

$$\begin{aligned} S_x^2 &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \end{aligned} \tag{8}$$

となり、「分散＝二乗の平均引く平均の二乗」と記憶しやすい。

例題

- **データ** $x = \{3, 7, 8, 12, 20\}$

$$\bar{x} = 10$$

-

$$\begin{aligned} S_x^2 &= \frac{1}{4}((3-10)^2 + (7-10)^2 + (8-10)^2 + (12-10)^2 + (20-10)^2) \\ &= 41.5 \end{aligned}$$

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

練習問題

- $y = 2x$ **とする。** y **の平均と分散を計算しなさい。**

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

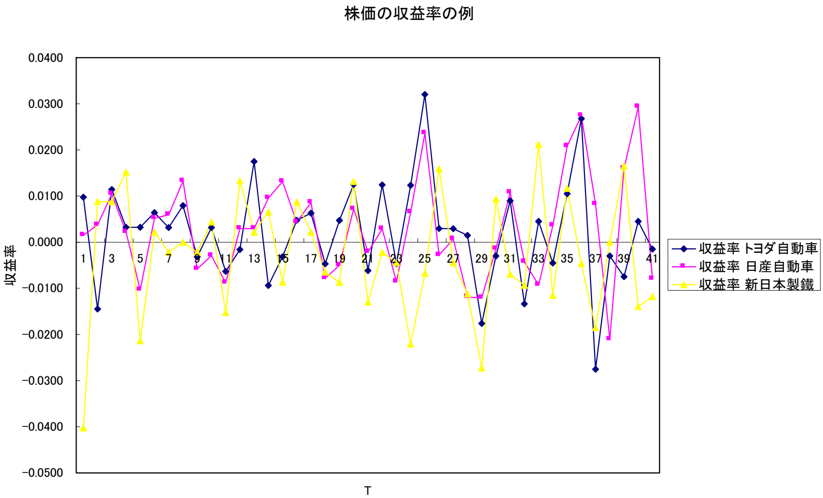
$$S_x = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- 標本標準偏差は分散と同様にデータの広がり（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, 0.0003), (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日産自動車である。株価の収益率の標準偏差がボラティィリティと呼ばれ、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）



まとめ

データの中心を表す代表値 平均、中央値、最頻値

データの広がり（散らばり）を示す代表値 標本分散、標本標準偏差、レンジ、平均偏差

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

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

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

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）

- $x = \{3, 7, 8, 12, 20\}$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$-2\times$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$-3\times$標本標準偏差，平均$+3\times$標本標準偏差）

練習問題

- $y = 2x$とする。yの2シグマ区間を計算しなさい。何個のデータが入っているのかを確認しなさい。

データの標準化

- $X = \{x_1, x_2, \cdots, x_n\}$ を標準化するというのは、X のデータから平均を引いて、標準偏差で割る操作である。標準化されたデータを Z とする $Z = \{z_1, z_2, \cdots, z_n\}$。

$$z_1 = \frac{x_1 - \overline{x}}{s_x}, \quad \cdots, \quad z_n = \frac{x_n - \overline{x}}{s_x} \tag{9}$$

- 偏差値：偏差値は標準化変化の一種、先の Z を用いれば偏差値 $= 50 + 10z$
- 標準化されたデータの平均は0、分散が1となる

証明

- **標準化されたデータの平均は** 0、**分散が** 1 **となる**

Proof.

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{10}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \overline{x}}{s_x} \tag{11}$$

$$= \frac{1}{s_x}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x}\right) \tag{12}$$

$$= 0 \tag{13}$$

□

続き

Proof.

$$S_z^2 = \frac{1}{n-1}\sum_{i=1}^{n} z_i^2 \tag{14}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^2 \tag{15}$$

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

$$= 1$$

□

練習問題

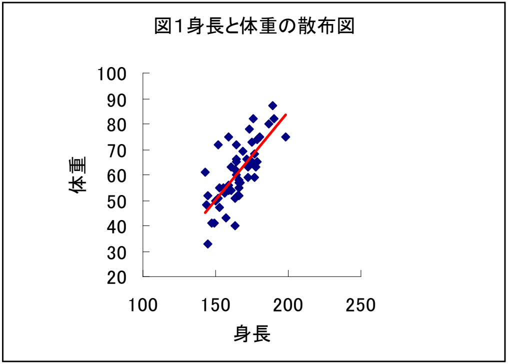
$x = \{2, 4, 5, 5\}$ に関して標準化しなさい、得られた新たな変数を z とする。z の平均と分散を計算しなさい。

二変数の間の関係、散布図と相関係数

- 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。

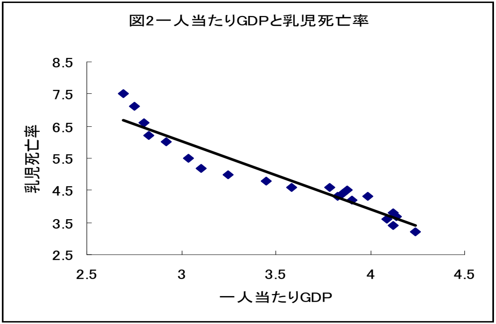
散布図

- 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

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

- 共分散の分解

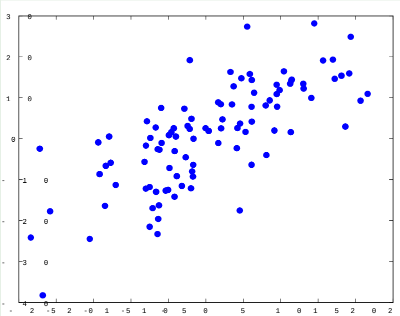
$$S_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

手計算では分解公式のほうが簡単。

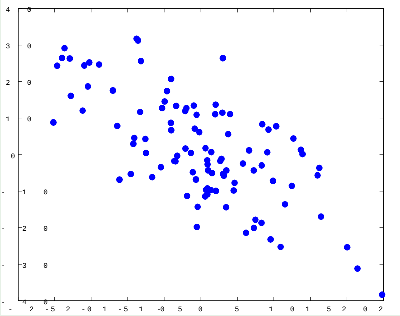
標本共分散（続き）

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。

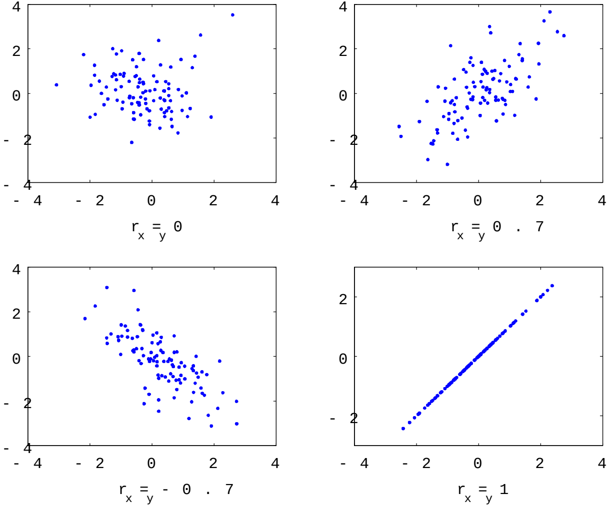
$$
\begin{aligned}
r_{xy} &= \frac{S_{xy}}{S_x S_y} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\b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

- **一つ重要な性質：任意の二つの変数の相関係数は**

$$-1 \leq r_{xy} \leq 1$$

である。相関係数が -1 と $+1$ にはさまれた数値で表現できるので実用上便利、比較しやすい。

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

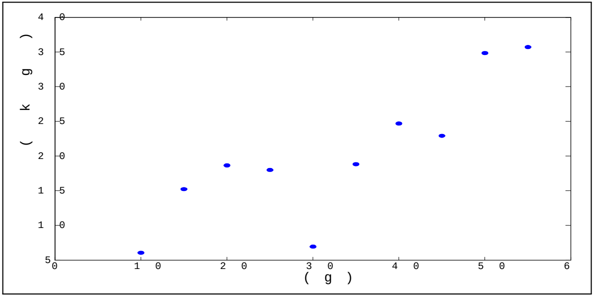


図5 肥料と農産物の出来高との関係

相関係数を計算したら、$r_{xy} = 0.84$ 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。

練習問題 $x = \{4, -3, 5, 1, 5\}, y = \{1, -3, 3, 0, 1\}$ とする。x と y の平均、分散、共分散及び相関係数を計算してください。