データの分析【統計的探究プロセス・回帰分析】　p.192～193

1 学習内容の説明 ⇒ 2 問題演習 ⇒ 3 振り返り（確認テスト・相互採点・リフレクションの記入）

【内容目標】統計的探究プロセスはいろいろな場面で使えるようになろう

研究　統計的探究プロセス

実社会では，さまざまな社会的問題に応じて，統計的手法を用いた問題解決が行われている。そのときには，

「問題 → 計画 → データ → 分析 → 結論」

の5段階からなる **統計的探究プロセス** を意識することが大事である。

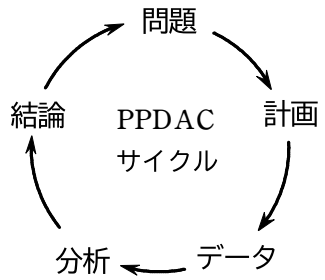
問　題（Problem）　… 解決すべき事柄を把握し，統計で扱える問題を設定する。

計　画（Plan）　… 設定した問題に対して，集めるべきデータと集め方を考える。

データ（Data）　… 計画にしたがってデータを集め，表などに整理する。

分　析（Analysis）　… 目的やデータの種類に応じてグラフにまとめたり，データに関する数値を求めたりして，特徴や傾向を把握する。

結　論（Conclusion）　… 見いだした特徴や傾向から結論をまとめて表現したり，さらなる課題や改善点を見いだしたりする。



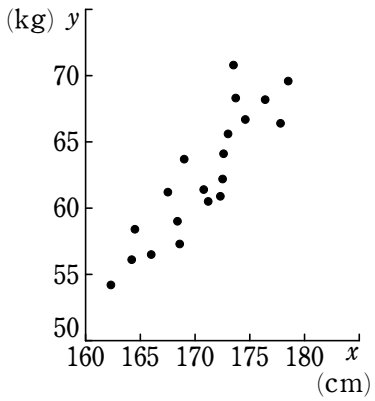
また，実社会でのデータは，一般に非常に大量であり，手計算では処理しきれないことがほとんどである。そのような大量のデータを扱う際には，コンピュータなどの情報機器を用いて，グラフをかいたり，さまざまな計算を行うとよい。

コラム

回帰分析

ある高校の1年生男子の身長 x と体重 y の散布図について，これらの点は，ある直線の近くに並んでいるようにも見える。

そこで，このデータの傾向を最もよく表す1次関数を見つけることを考えよう。



散布図において，点の配列に「できるだけ合うように引いた直線」を**回帰直線**という。そこで，この回帰直線をこの散布図の中に引くことを考える。

データの分析【統計的探究プロセス・回帰分析】　p.192～193

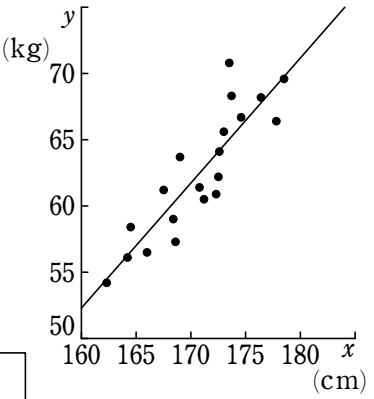
直線を引く基本的な方法は回帰分析と呼ばれるもので，コンピュータなどの情報機器を利用してかくこともできる。

実際には手計算で回帰直線を引くことは少なく、
表計算ソフトなどを利用することがほとんどである。

上の x と y のデータでは，次の 1 次関数が得られる。

$y=0.943x-98.603$

実際に直線を引くと右の図のようになる。



回帰分析は，自然科学のデータ分析で必須であるだけでなく，経済学や社会学などの社会科学を学ぶ上でも重要な手法である。

回帰直線を引くことで、１つの変量のデータから
もう一方の変量のデータの値を予測することが可能である。

研究　最小２乗法

実際のデータはある程度散らばっているため、どの部分に直線を引くのが妥当かが問題となる。その解決方法の一つとして「最小２乗法」がある。これはガウスによって体系的に理論化され、未知の真の値を観測によって推定する方法として有効であるとされている。最小 2 乗法は，人工知能（AI）の分野でも用いられる。

データ d_1，d_2，d_3，d_4，d_5 が与えられたとき，データとの差の 2 乗の和

$(d_1-x)^2+(d_2-x)^2+(d_3-x)^2+(d_4-x)^2+(d_5-x)^2$　……①

を最小にする x の値について考えてみよう。① を x の関数とみて $f(x)$ とおくと

$$f(x)=(d_1-x)^2+(d_2-x)^2+(d_3-x)^2+(d_4-x)^2+(d_5-x)^2$$
$$=5x^2-2(d_1+d_2+d_3+d_4+d_5)x+{d_1}^2+{d_2}^2+{d_3}^2+{d_4}^2+{d_5}^2$$

ここで，　$a=\dfrac{d_1+d_2+d_3+d_4+d_5}{5}$　（データの平均値）

$b=\dfrac{{d_1}^2+{d_2}^2+{d_3}^2+{d_4}^2+{d_5}^2}{5}$　（データの 2 乗の平均値）

とおくと　$f(x)=5x^2-2\cdot5a\cdot x+5b=5(x-a)^2+5(b-a^2)$

よって，$f(x)$ は x の 2 次関数で，$x=a$ で最小値 $5(b-a^2)$ をとる。

ここで，最小値を与える x の値 a はデータの平均値であり，最小値をデータの大きさ 5 で割って得られる $b-a^2$ はデータの分散である。

データとの差の 2 乗の和を最小にするものを求めることは，観測値から未知の真の値を推定する方法として広く用いられている。これを **最小 2 乗法** という。

また、このように、最小 2 乗法によってデータに最もよくあてはまる直線を求めることを、**線形回帰**（linear regression）といい，得られた直線を**回帰直線**という。

最小 2 乗法で求められた直線の方程式の係数は，平均値，標準偏差，相関係数で表される。つまり，2 つの変量 x, y に対して最小 2 乗法による回帰直線の方程式は

$$y-\overline{y}=r\cdot\frac{s_y}{s_x}(x-\overline{x})$$

$\overline{x}$，$\overline{y}$ ：それぞれの x，y の平均値　s_x，s_y ：それぞれの x，y の標準偏差　r ： x と y の相関係数

となり，回帰直線は点 $(\overline{x},\ \overline{y})$ を通ることがわかる。