第 9 章

相関と線形回帰

得られたデータから何らかの傾向を見つけ出すことは，統計的なデータ処理における重要な目的のひとつである．現実には散布図を見ただけでも分かるような傾向もあれば，判定に困る微妙なケースも少なくない．そのようなときに客観性をもった判断を行うにはどうしたらよいのだろうか．本章では，そのことについて考えてみる．

9.1　データの相関

一般に背が高い人ほど体重は重い．また，所得が多い人ほど，その人が住む住宅の面積も概して大きいにちがいない．一方，次のような例も考えられる． — 世界の国や地域を比較した場合，一人当たりの台所洗剤の使用量が多い地域ほどガンによる死亡率は高い．— このような事例は，自然現象あるいは人間や社会に関わる現象の中にいくらでも探し出すことができる．

そして以上の例のように，2 種類の数値データが互いに何らかの関連をもっている場合，データの間に**相関** (correlation) があるという．本章では数値データの相関について統計的に分析する手法を扱う．

9.1.1　相関と因果関係

冒頭に挙げた例のうち，身長と体重の相関や所得と住居面積の相関は，直接的な原因と結果の関係から導かれる．なぜなら，身長が伸びるときには体が大きくなるのだから，当然体重も増加するはずだし，お金持ちは家に大きな費用を支出できるからこそ，広い住居に住むことになる．

しかし，地域の間の比較で，中性洗剤の使用量とガンの発生率について単純な原因と結果があるということはできない．このような傾向が見られることは，むしろ次のように解釈するのが適切である． — 一般に低開発国では中性洗剤の利用は少ない．工業製品が所

得水準にくらべて割高で，低所得の住民には手がとどかないからである．一方，低開発国ほど平均寿命は短く，高齢者の割合が先進国に比べて少ないのである．そのため人口比でみたガンの発生率は低い[*1]．したがって，中性洗剤の使用量とガンの発生率の間に相関が現われることはあっても，それは中性洗剤がガンを引き起こしているというわけではないのである．

2 つの現象のうち一方が他方の結果になっているとき，これらの間には**因果関係 (causality)** があるという．2 つのデータの間に因果関係があれば，なんらかの相関が表れると考えてよい．ただし，その他の要因が加わることで相関が分かりにくくなることもある．

因果関係があれば相関が生じることが多い

複数のデータの間に相関があれば，なんらかの因果関係の存在が示唆される．しかし，相関があっても因果関係は存在しないことも珍しくない．上のガンと洗剤のケースなども，容易に「中性洗剤はガンの原因になる」などというインチキな話にすりかえることができる．このように相関関係を因果関係に結びつける論法は，統計で人をだます誤った推論の代表的なものである．

相関があれば因果関係があるかも知れない．
しかし，必ずしもそうとは限らない

9.1.2　散布図 — 相関をグラフで見る

2 種類のデータの相関を見るために，横軸と縦軸を使ってデータをプロットしたものを**散布図 (scatter diagram)** という．散布図を使うと直観的に相関の有無を判断することができる．

図 9.1 に示された例は，$\boldsymbol{X}, \boldsymbol{Y}$2 系列のデータの間に相関がないものから，完全な相関をもつものまで，4 つの場合について散布図を示した．この図にあるように，**正の相関**はデータ X が増加するときに Y も増加するような相関，**負の相関**はデータ X が増加するときに Y が減少するような相関である．

図 9.1(2) のように弱い相関が見られる場合というのは，それが何らかの必然性によるものであるかどうかの判断は困難である．つまり特に理由がなくても，データのばらつきによって**偽の相関**が生じる確率は無視できない．この問題については後に詳しく議論す

[*1] WHO の資料で見ると，先進国においてはガンが主要な死因のひとつであり，その死者の割合は死者全体の 12.7% を占めるのに対し，低開発国においては，感染症や出産時の死亡，エイズなどが主要な死因を占めていて，ガンによる死者の割合はきわめて低い (2005 年のデータ)．
`http://www.who.int/mediacentre/factsheets/fs310/en/`

る．しかし (3)(4) のように，明らかに傾向が読み取れるようなケースでは，なんらかの必然的な原因があって，相関が現われていると考えてよい．ただし，繰り返しになるが，この場合でも因果関係があると即断してはいけない．

これらの図の欄外に示されている ρ_{xy} という数値は**相関係数**と呼ばれるもので，相関の強さを表す重要な統計量である．これについては次節で詳しく解説する．

なお，ここでは 2 系列のデータの間に隠れている相関は直線的な関係，つまり 1 次関数で表されるような単純なものであると仮定する．このような相関を 1 次または**線形の相関**という．この章では主に線形の相関について考えることにする．

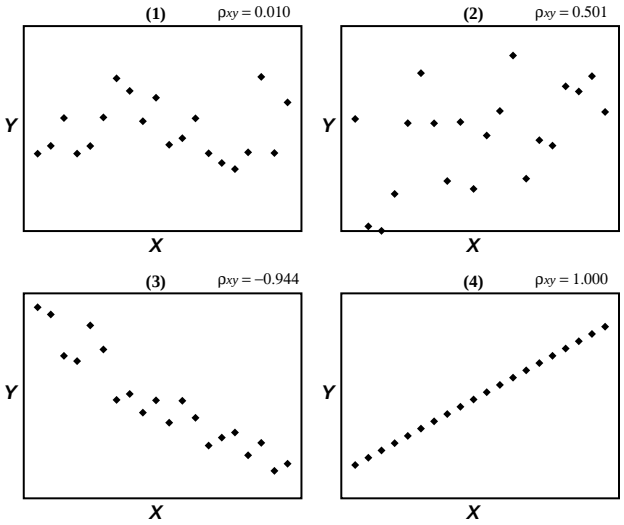

図 9.1　散布図で示された相関の有無: **(1)** 相関なし **(2)** 弱い正の相関 **(3)** 強い負の相関 **(4)** 完全な正の相関

9.2　相関係数と線形回帰

9.2.1　共分散

たとえば国ごとの人口密度と出生率とか，人ひとりずつの喫煙年数とガンによる死亡率，各自治体の教育予算が予算全体に占める割合と大学進学率のように，サンプルごとに 2 つの変数を知ることができるものとする．これらのデータを次のように 2 つの変数列 $\boldsymbol{x}$, $\boldsymbol{y}$ で表すことにする．

$$\boldsymbol{x} = x_1, x_2, x_3, \ldots, x_n$$
$$\boldsymbol{y} = y_1, y_2, y_3, \ldots, y_n$$

さて，次の式 (9.1) の定義で表される σ_{xy} は，$\boldsymbol{x}, \boldsymbol{y}$ の**共分散 (covariance)** と呼ばれる．

$$\begin{aligned} \sigma_{xy} &= \frac{1}{n}\sum_{i=1}^{n} \delta x_i \delta y_i \\ &= \frac{1}{n}(\delta x_1 \delta y_1 + \delta x_2 \delta y_2 + \ldots + \delta x_n \delta y_n) \end{aligned} \tag{9.1}$$

ここで δx_i は x_i の偏差 $x_i - \overline{x}$ を意味し，他も同様である (p.6)．

式 (9.1) は次の形をしていることに注目しよう．

共分散 ＝ （x_i の偏差 × y_i の偏差）の平均

共分散の定義は，1 章で出てきた p.7 の分散の定義を拡張したものであり，式 (9.1) で $x_i = y_i$ と置けば，そのまま分散の式が得られる．

例題 9–1　共分散の定義の式で y_i を x_i に置き換えると，分散の定義の式になることを確かめなさい (解答は省略)．

なお，共分散 σ_{xy} は次のように変形され，計算に当たってはこちらのほうが使いやすい．この形は，**積の平均 − 平均の積** が共分散になることを示している．

$$\sigma_{xy} = \frac{1}{n}\sum x_i y_i - \overline{x}\ \overline{y} \tag{9.2}$$

9.2.2　共分散の意味

ここで共分散の意味を考えるために，散布図上で 4 個のデータ点 S_1, S_2, S_3, S_4 が図 9.2 のように長方形の形に分布している状況をまず考えてみよう．この図で S_1 と S_2 の

データの平均値は，x が増加しても増加していない．すなわち，このように長方形の各頂点に 4 点があるようなデータでは，相関はない．

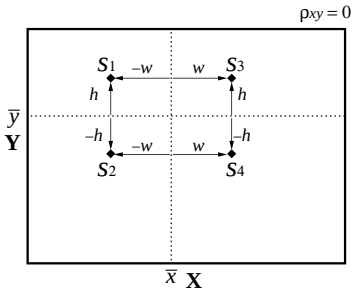

図 9.2　共分散がゼロになるデータの配置．2 つの点線は $\boldsymbol{x}$ と $\boldsymbol{y}$ の平均値を表す．

さて x_i, y_i の偏差というのは，x_i, y_i がそれぞれ平均値からどれだけずれているかを示す値のことであるから，図 9.2 のようなデータの配置では，x, y の偏差はいずれも図のように絶対値が等しい．そのことから，式 9.1 の値がどうなるかを考えてみよう．

$$\sigma_{xy} = \frac{1}{4}\{(-w)\times h + (-w)\times(-h) + w\times h + w\times(-h)\} = 0$$

すなわち共分散 σ_{xy} はゼロになる．

もっと一般の場合には，共分散はどのような値をとるのだろうか．図 9.3 の散布図をみてほしい．ここでは多数のデータがランダムに，つまり相関がないようにして与えてある．このとき，共分散はどのような値をとるかを考えてみよう．

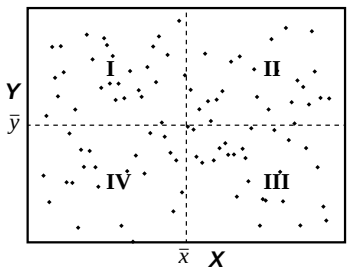

図 9.3　相関のないランダムなデータの散布図．破線は $\boldsymbol{x}$ および $\boldsymbol{y}$ の平均値．

まず $\boldsymbol{x}$ の偏差は，領域 **II** と **III** では正，領域 **I** と **IV** では負である．また $\boldsymbol{y}$ の偏差は，領域 **I** と **II** では正，領域 **III** と **IV** では負になる．したがって，式 (9.1) に現われる偏差の積 $(x_i - \overline{x})(y_i - \overline{y})$ の値は，領域 **II** と **IV** では正に，領域 **I** と **III** では負になることがすぐに分かる．

さて，これらのデータはランダムに与えられているのであるから，**I, II, III, IV** どの領域にも同じ確率で出現し，また図の破線の両側に対称に現われると考えてよいだろうから，4 つの領域のデータの偏差の積を足し合わせたものは，期待値がゼロになるはずである．すなわち，**相関をもたないランダムなデータの場合には，共分散の期待値はゼロになる．**

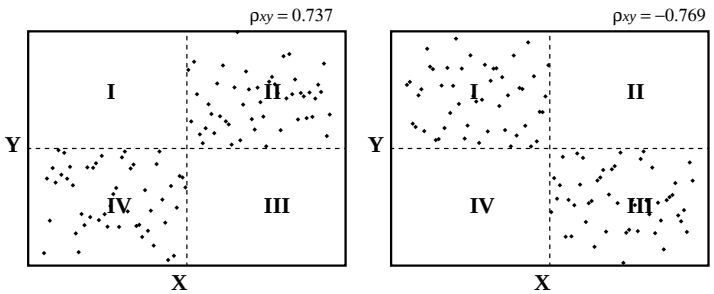

図 9.4　正と負の相関が生じる分布

もしも図 9.4 のようにデータが領域 **II** と **IV** にだけあったとしたら，結果はどうなるだろうか．今度は偏差の積が正の部分だけしかないのだから，その和は正，結局共分散は正になる．このとき，データは右上がりに分布しているから，結局，**データが右上がりに分布しているときには共分散の期待値は正になる．**逆に，**データが右下がりに分布しているときには共分散の期待値は負になる．**

9.2.3　相関係数

ある変数ともうひとつの変数の間の相関がどの程度強いかを表す量として**相関係数 (correlation constant)** がよく用いられる．相関係数は ρ_{xy} で表されることが多く[*2]，式 (9.3) で定義される．

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{9.3}$$

ここで，σ_x，σ_y はそれぞれ変数 $\boldsymbol{x}$, $\boldsymbol{y}$ の標準偏差であり，σ_{xy} は前節で登場した共分散である．この定義と前節での考察から，相関係数の値は，相関がなければゼロに近い値になることがわかる．それでは相関が強いときには相関係数はどのようになるのだろうか．

最も強い相関は図 9.1(4) のように，散布図でデータが直線状に配置されている状態である．ただし，直線が水平または垂直の場合には，相関がないことに注意しよう．つまり，強い相関というのは式 (9.4) のように y_i が x_i の 1 次式で表される状況である．

$$y_i = ax_i + b, \quad (i = 1, 2, \ldots, n) \tag{9.4}$$

[*2] 他に r_{xy} もよく使われる．なお，ρ は「ロー」と読む．

計算を簡単にするために $b=0$ として，次の比例関係が $\boldsymbol{x}$ と $\boldsymbol{y}$ の間にあるものとしよう．

$$y_i = ax_i,\ \ (i=1,2,\ldots,n) \tag{9.5}$$

このとき共分散 σ_{xy} を求めてみよう．まず δy_i は次のように変形される．

$$\delta y_i = y_i - \overline{y} = ax_i - a\overline{x} = a\delta x_i \tag{9.6}$$

ここですべての y_i が a 倍されると，それらの平均も a 倍されることを使っている．

式 (9.1) にこれを代入すると，

$$\begin{aligned}\sigma_{xy} &= \frac{1}{n}\sum_{i=1}^{n}\delta x_i\delta y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\delta x_i \times a\delta x_i \\ &= a\frac{1}{n}\sum_{i=1}^{n}\delta x_i^2 = a\sigma_x^2\end{aligned} \tag{9.7}$$

次に分散 σ_y^2 について考えてみよう．

$$\begin{aligned}\sigma_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(\delta y_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(a\delta x_i)^2 = a^2\frac{1}{n}\sum_{i=1}^{n}(\delta x_i)^2 \\ &= a^2\sigma_x^2\end{aligned} \tag{9.8}$$

これから $\sigma_y = \sqrt{\sigma_y^2} = |a|\sigma_x$ となり，相関係数は次のように表されることになる．

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{a\sigma_x^2}{|a|\sigma_x^2} = \pm 1,\ \ (a>0 \text{ のとき } 1,\ a<0 \text{ のとき } -1) \tag{9.9}$$

結局，式 (9.5) のように y_i が x_i に比例するときには，傾きの正負に対応して相関係数は 1 か -1 になることが分かった．この結論は，前節で，相関がないときには共分散の期待値がゼロになり，その結果相関係数の期待値もゼロになることを確かめているので，まとめると，2 つのデータ系列 $\boldsymbol{x}$, $\boldsymbol{y}$ の相関係数 ρ_{xy} とデータの相関の間には次のような関係がある．

相関係数 $=\pm 1$　$y_i = ax_i + b$ で表されるような完全な相関があるとき
相関係数 ≈ 0　相関がないとき

一般には相関係数はこれらの中間的な値をとり，0.9 以上であれば**強い相関**があるといえる．相関がないときの相関係数の 期待値 はゼロであるが，あるデータ列から得られる相

関係数は一般にゼロを中心としたある種の分布を作る．そのため，たとえば相関係数が 0.4 であった場合に，相関があるという判断を単純に行うことは無理がある．その問題については 9.2.5 節で検討する．

9.2.4 線形回帰 (最小二乗法)

相関のあるデータを散布図にして，その中にデータの変化の傾向を表す直線を引きたいことはよくある．なお，このように全体の傾向を表す関数のことを**モデル**と呼ぶこともある[*3]．

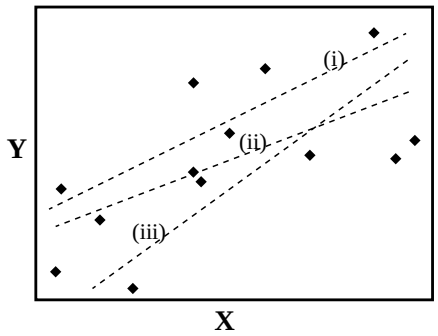

図 9.5 すべてのデータ点を代表する直線はどれがよいか

図 9.5 に描かれた 3 本の直線のうち，適切なものが (ii) であることは，勘で分かる．しかし客観的に最良のモデルとなる直線を決定するにはどうしたらよいだろうか．そのために用いられるのが**線形回帰 (linear regression)** または**最小二乗法** (**least-square method**) と呼ばれる方法である．

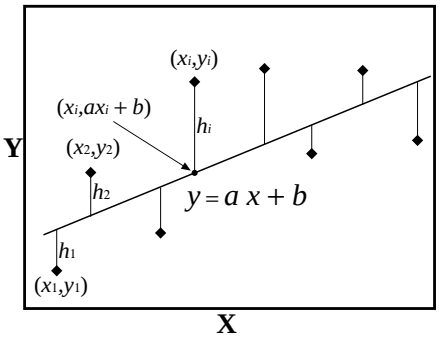

図 9.6 最小 2 乗法の原理: $h_1^2 + h_2^2 + \ldots$ が最小になるように a, b の値を決めてやる．

[*3] モデルという意味は次のようなことだ．— 与えられたデータ点の中に潜む関係としてはいろいろなものが考えられる．その中で単純な線形な関係を，仮にひとつの「モデル」として仮定してデータをそれに当てはめることで，問題を分析したり解釈したりしたい．

図 9.6 のようにデータ点 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ が与えられていたとき，$y = ax + b$ で表される直線を引いたとしよう．このとき，i 番目のデータ点と直線の縦のずれを h_i とすると，

$$h_i = y_i - (ax_i + b) \tag{9.10}$$

となる．直線 $y = ax + b$ は，a と b の値を変化させることで，傾きを変えたり平行にずらしたりできる．それでは a, b がどのような値をとったときに，直線はデータ点をもっともよく近似できるだろうか．詳しい導き方は付録（p.153）にゆずり，ここでは結果だけを示す．

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{9.11}$$

$$b = \overline{y} - a\overline{x} \tag{9.12}$$

式 (9.11),(9.12) で得られた a, b を用いて直線を引くと，データの増減を「ほどよく」表した直線が得られる．このような直線を**回帰直線**と呼ぶ．

例題 9–2　成人男子 6 人の靴のサイズと身長を調べたところ，次のようなデータの組が得られた．これらを散布図にプロットし，最小 2 乗法を使って回帰直線の係数を求めて，図に直線を描きなさい．

$(24.5, 165.4), (28.0, 182.7), (26.0, 171.6), (25.5, 173.1), (25.0, 175.1),$
$(24.0, 170.6)$

これらのデータの組を $(x_1, y_1), (x_2, y_2), \ldots$ とすると，式 (1.6) から分散 σ_x^2 が，式 (9.2) から共分散 σ_{xy} が得られる．具体的には x_i，y_i，x_i^2，y_i^2，$x_i y_i$ それぞれの平均，$\bar{x}, \bar{y}, \bar{x^2}, \bar{y^2}, \bar{xy}$ を個別に計算してから分散と共分散を求め，式 (9.12) に代入すればよい．電卓を使ってもかなりめんどうなので，Excel などを利用するとよい．

データの散布図と，このようにして求めた a, b を使って引いた直線を図 9.7 に示した．

以上のようにして求められた回帰直線で示される相関が，真の相関であるのか，あるいは母集団には相関がないのに，抽出によってたまたまある傾向が現れてしまったのかという疑問は，特にデータの点が少ないときには問題になる．このことについては次節で考えてみよう．

9.2.5　相関の有無を検定する

前述のように，母集団が全く相関をもたない場合でも，そこから無作為抽出を行った場合の標本の相関係数は一般にはゼロにならず，ある範囲で分布する．図 9.8 は，ランダム

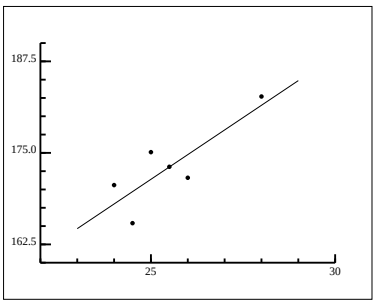

図 9.7　6 組のデータから得られた回帰直線

に 10 個の点を発生させて，その相関係数を計算したものである．真の相関は存在しないはずなのに，場合によってはかなり大きな相関係数が出現してしまうことがわかる．

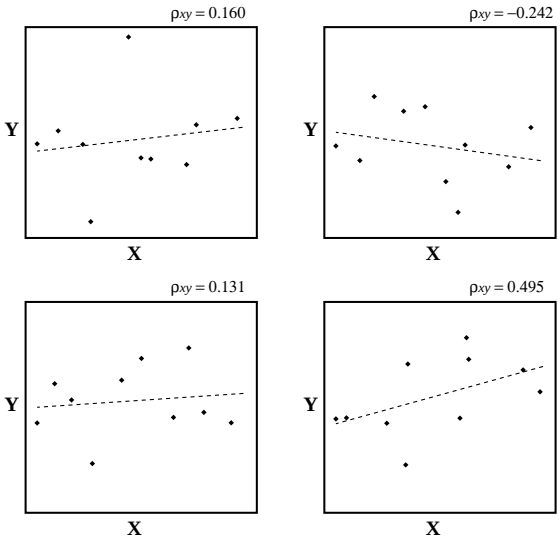

図 9.8　ランダムな操作で作られた偽の相関

この例のように，真の相関はないのに現われてしまう**偽の相関**を何らかの方法で検定して排除するにはどうしたらよいだろうか．ここでは次の定理を使うと，一定の仮定の下に相関の有無を検定することができる．

[定理] 相関を持たない母集団があり，ただしその母集団が **2** 次元の正規分布 をしているとしたとき，そこから大きさ n の標本を無作為抽出したとすると，標本の相関係数 ρ_{xy} について，次の式 (**9.13**) のように定義される量 T は自由度 $n-2$ の t 分布に従う．

$$T = \sqrt{\frac{(n-2)\rho_{xy}^2}{1-\rho_{xy}^2}} \tag{9.13}$$

ここで 2 次元の正規分布というのは，図 9.9 のように 2 つの量がそれぞれ正規分布しているようなものをいう．図のように，2 つの量の間には相関がある場合もない場合もある．上の定理では，図 9.9 の左のように相関が全くない母集団から n 個の標本を抽出して散布図を作って相関係数 ρ_{xy} を求めることを想定している．この場合，何度も抽出を繰り返すと，その度に ρ_{xy} は異なった値をとるが，T のような量を計算してやると，その値は自由度 $n-2$ の t-分布をつくるというわけだ．

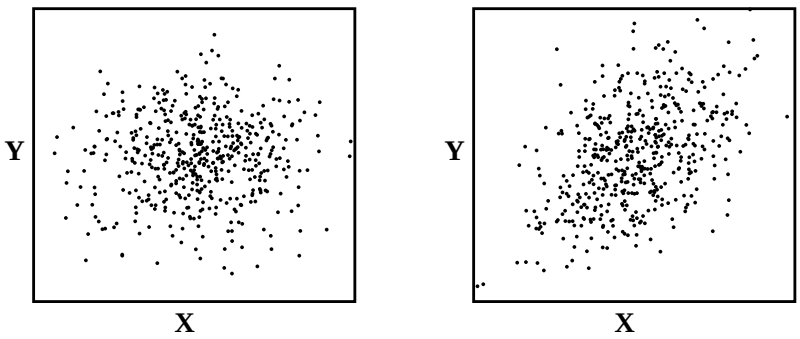

図 9.9　二次元の正規分布をもつデータの散布図：左は相関なし，右には正の相関がある．

ここで 9.2.4 節で出てきた例題をふたたび考えてみよう．図 9.7 を見ると，データ点は回帰直線からかなり離れていて，しかも標本数は 6 しかない．このとき，本当に相関があるのかどうかは，微妙な問題である．そこで上の定理を用いて検定してみよう．

例題 9–3　成人男子 6 人の靴のサイズと身長を調べたところ，次のようなデータの組が得られた．このデータから，靴のサイズと身長には相関があるといえるかどうかを検定しなさい．

$(24.5, 165.4)$, $(28.0, 182.7)$, $(26.0, 171.6)$, $(25.5, 173.1)$, $(25.0, 175.1)$,
$(24.0, 170.6)$

ここでは，靴のサイズと身長には相関がないという仮説，つまり帰無仮説を立てて，それが棄却できるかどうかを検定する．

式 (9.3) を用いれば，このデータにもとづく相関係数は，最小 2 乗法の計算のついでぐらいの手間で求めることができ，実際に計算してみると，その値は $\rho_{xy} = 0.8323$ になる．なお，ここではすべての成人男子を母集団として，そちらのほうの相関係数は ρ_{xy} で表しているので，異なった表記法を使っている．

これから，式 (9.13) を使って，T を計算してみると，$n = 6$ であるから，

$$T = \sqrt{\frac{(6-2)0.8323^2}{1-0.8232^2}} = 3.00$$

となる．ここで t-分布表を見ると，自由度 4 のとき，3.00 という値は $\alpha = 0.01$ と $\alpha = 0.025$ の間に来る．つまり，母集団に相関がないと仮定した場合には，T は 0.025 以下の確率でしか現われないはずである．よって，靴のサイズと身長に相関がないという仮説は危険率 0.025 で棄却される．すなわち相関はある．

ここで注意しておくと，T は負の値をとることはないので，片側検定と同じく分布の正の領域だけで判断すればよい．つまり両側検定のときのように $\alpha = 0.025$ から危険率を 0.05 とすることはない．

現実にデータの間に相関があるかないかを判定したいことは，よくあることだ．散布図を見て感覚的に判定することもあるが，このように「きわどい」データの場合には判定に苦しむことになる．前提として母集団に正規分布が仮定できるというときという制限はあるものの，この検定はそのような場合に客観的な判定法を与えてくれて，実用的な意味が大きい．

9.2.6　線形でない相関

ここまで見てきたように，x_i と y_i が式 (9.4) のような 1 次式の関係にあるときは，線形の相関，あるいは 1 次の相関があるという．しかし，2 つの変数の間のの相関が直線的でない場合もしばしばある．たとえば図 9.10(a) のように変数の関係が指数関数に沿って分布している場合や，図 9.10(b) のように，2 次曲線に沿って分布している場合では，相関係数を求めることは意味がない．図 9.10(b) のような場合には，明確に相関があるにも関わらず相関係数がゼロに近くなってしまう．

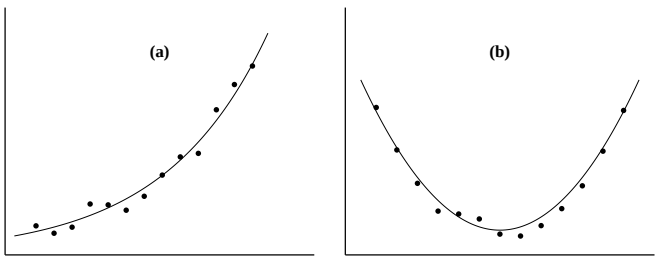

図 9.10　非線形の相関の例： (a) 指数関数に近似できる相関 (b)2 次関数に近似できる相関

このような場合には，**非線形最小 2 乗法**を使うことがある．具体的な手法についてはここでは述べないが，要点だけ解説しておこう．

たとえば，あるデータの傾向が理論モデルにもとづく予測から 2 次関数に従うと予測されているとしよう．だれでも一度は勉強したように，2 次関数は一般に次の形をしている．

$$y = ax^2 + bx + c$$

ここで係数 a, b, c を適当に与えることによって，最小 2 乗法のときと同じように，データと理論曲線の間の偏差から求められた 2 乗和を最小にするように工夫すればよい．このような手続きで，データをある与えられた関数で近似することができる．当然のことながら，これは直線でなく曲線なので，**回帰曲線**という．

観測されたデータがどのような関数の形に従うのかを決定することは，法則性を調べたり因果関係を推定する上で重要な意味をもつことが多い．そのような場合に，非線形の最小 2 乗法を用いて回帰曲線を求めることは，有力な解析の手段になり得る．

【章末問題】

問題 9–1 たまたま大きな卵を割ってみたときに，黄身がそれほど大きくないように思った A さんは，8 個の卵を買ってきて，全体の質量 (x g) と黄身の質量 (y g) を調べてみた．その結果，次のような結果を得た．

全体 x	62.2	42.8	61.8	79.3	63.1	51.4	60.9	69.9
黄身 y	36.7	28.7	32.0	37.7	31.8	31.5	32.3	34.8

1. x, y それぞれの分散と x と y の共分散を求めなさい．
2. x, y の相関係数 ρ_{xy} を求めなさい．
3. このデータを $y = ax + b$ に回帰したときの係数 a, b を求めなさい．
4. 式 (9.13) を使って， T の値を計算しなさい．さらに卵の質量と黄身の質量の間には相関がないという仮説を，危険率 0.01 で検定しなさい．