第１０章　信頼区間推定

～誤った結論を導かないための～

電波の伝搬など自然を対象にする不規則現象は第 1 部で述べた確率分布に支配されるが、その結果は統計的性質として現われてくる。そのため、測定データの評価や無線通信の回線設計を高信頼なものにするためには、統計的なものの見方が大事になる。本章では、多岐にわたる統計の分野から、少数データから、あるいはばらつきの大きいデータから誤った結論を導かないための信頼区間推定を取り上げる。

注：本章 10.1～10.4 の部分は[1]の 4.1 節の記述を転載している

10.1　信頼区間推定の概要

物事には原因があって結果が起きる。原因となる物理量の変数（説明変数、あるいは基準変数）の値を x、その観測量の変数（目的変数）を y とし、x と y には次式で与えられる線形関係が有るとする。統計で言うところの**母集団**の特性である。

$$y(x) = a + bx \tag{10.1}$$

実際には、これに誤差成分（e）が加わり、観測される目的変数の試行 i 毎の値は次式となる。

$$y_i = a + bx_i + e_i \tag{10.2}$$

N 個のデータの処理によって得られる**回帰直線**

$$\hat{y}(x) = \hat{a} + \hat{b}x \tag{10.3}$$

は、母集団の特性、すなわち、(10.1)式の y を推定するものになる。この(10.3)式で表されるモデルは、**単回帰モデル**と呼ばれる。単回帰モデルでは、通常、以下の前提が採用される[2]。

i) 変数 x のサンプル値 x_i は誤差を含まない。（誤差は y_i のみ（グラフの縦軸方向のみ））

ii) 誤差は毎回独立である（独立性）

iii) 誤差の期待値は 0 である（普遍性）

iv) 誤差の大きさ（標準偏差）は x の値によらない（等分散性）

v) 誤差の分布は正規分布に従う（正規性）

回帰直線の決定にはガウスが編み出した**最小二乗法**が用いられる。最小二乗法による推定は**最尤推定**になる。最小二乗法では回帰直線と観測値とのずれ $\varepsilon_i = y_i - \hat{y}_i$ の２乗和（平均二乗残差）が最小になるよう $\hat{a},\hat{b}$ を定める。回帰直線は、データが与えられれば一意に定められる。この回帰直線は、傾向を捕らえるのに分かりやすいが、どの程度信用できるのかという不安がある。

回帰直線の信頼性を定量的に評価するために生み出された概念が**区間推定**である。本当の特性が有りそうな範囲を上限と下限で囲み、**信頼区間**と呼ぶ。この確率の目安値を 90%, 95%, 99%などで定め、95%が良く用いられる。95%信頼区間の場合、その範囲内にある全ての特性（直線で表される）について 5%以上の存在確率が有り、どれも棄却できない、とみなす（注１）。要は、「回帰直線だけで判断するのは危ないですよ、区間内に直線で引かれる傾向は、全部有りそうですよ」ということである。値を点で推定する最尤推定が、その傾向を積極的に見出そうとするものであるのに対して、幅で推定する区間推定は、判断に慎重さを求める確実性志向の推定になる。

具体的な信頼区間（上限と下限）の定め方は後に述べるが、ここでは、信頼区間の特徴を図 10.1 により説明する。真の特性（$y = 3+x$ を仮定）を点線①で、ばらつきのある 10 のデータ（●）に対する回帰直線を実細線②で、95%信頼区間の上限と下限を実太線③で示している。確率を 99%にすれば信頼度は上がるが、推定区間は広くなり傾向が見えにくくなる。逆に 90%にすれば、推定範囲は狭まるが、信頼性が落ちると言うジレンマがある。図の回帰直線は x の増加に対して y も大きくなる傾向を示しているが、例えば一点鎖線④のように、y が減少する特性も信頼区間の中に 5%以上の確率で存在しており、このデータからは「“x が大きくなれば y も大きくなる”と言う答えを出せない」が区間推定の約束事（ルール）になる。ただ、データ数が増えると信頼区間の幅が狭まり、x が大きくなれば y も大きくなるという結論が言えるようになる（後述する図 10.4）。繰り返しになるが、信頼区間推定は、傾向を積極的に見出そうとするものではなく、誤った判断をしないよう、だめなもの（＝信頼区間外の特性）を棄却するという用心深さに比重を置く推定になっている。

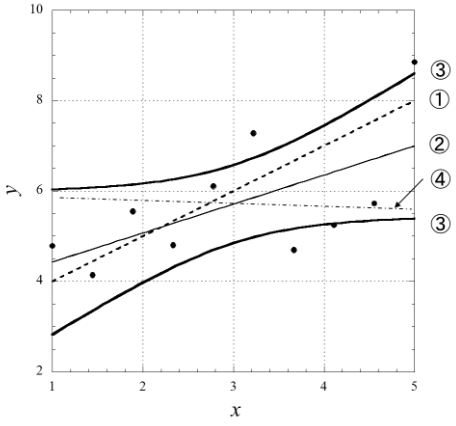


図 10.1　$y=3+x$ の特性①にばらつきを加えたデータ（●）の回帰直線②と 95%信頼区間の上限と下限③、その中に含まれる一つの例④（信頼区間はデータのばらつき幅を表しているものではないことに注意）

［注１］データから 95%信頼区間が定まったとき、「母平均が 95%の確率でその範囲に存在する」と考えるのは厳密には正しくない。なぜなら、母平均が固定値であるのに対して、範囲そのものはデータセット毎に変化し、その値は一回のデータセット限りであるからである。ゆえに、「100 回信頼区間を定めたら、そのうちの 95 回は確率の意味で母平均がこの範囲（毎回変わる）に存在する」がより正しい意味になる。

ティータイム　「回帰」の由来

本節の導入として回帰直線の概要説明を行った。回帰分析というような使われ方もする。この「回帰」の英語名 regression には、後退、後戻りと言うような意味がある。なぜこの言葉が使われるようになったのであろうか。これには、興味深い生い立ちがある。

イギリスの人類学者・統計学者・遺伝学者であったゴルトン（Francis Galton: 1822-1911）は、人の才能がほぼ遺伝によって受け継がれると主張する優生学を唱えたことで知られる。ゴルトンは父親の身長(x)と彼等の息子たちの身長(y)の統計分析を行い、身長の高い父親からは身長の高い子供ができる傾向が強いこと（すなわち x と y には正の相関があること）を調べた。例えば、両世代ともその平均値が 170cm であったとする。そこで、父親の身長が 180cm 近辺のグループの息子たちの平均身長を見ると、180cm より低い値に出た。すなわち人の遺伝においては、親の代が平均からずれた値であっても、次の代には平均値に戻ってゆく（＝後戻りしてゆく、回帰してゆく）性質があることを発見した。これが、回帰＝regression が用いられていることの由来のようである[2]。

この理屈を、相関を有する２系統の乱数で再現してみたい。図 10.2 は、相関係数が 0.7 である２系統（x と y）の正規乱数（平均 0、分散 1 の正規分布する乱数）の散布図である。y=x で示される点線の周りにばらつきを持って分布している。図中の楕円はデータの 90%が含まれるエリアを示しており、当然ながら、この楕円も点線を中心に 45°傾いている。このデータの回帰直線を最小二乗法で求めると、点線よりは傾きが小さい実線のように求められる。図より、平均からずれた x=2 を見ると、y の値は 1.5 程度に読める。これが、平均値（=0）への回帰（regression）の意味になる。どうも遺伝の仕組みと言うよりは、モデル化の考え方そのものに帰着する話と言えそう。不思議な感覚を持つと思うが、その興味を維持したまま、次項に進んでほしい。

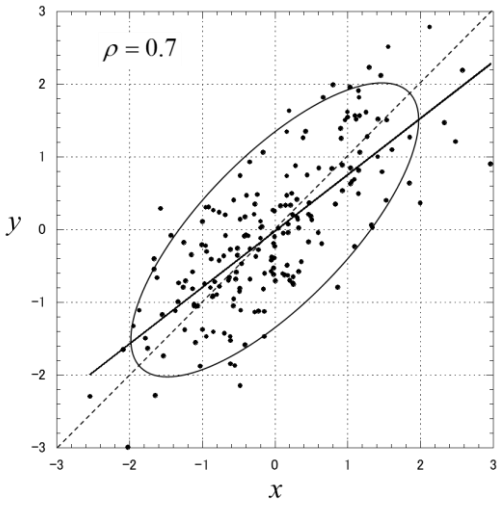


図 10.2　相関を有する二つの正規乱数の散布図と回帰直線（実線）

10.2　母平均の統計的推定

信頼区間を求める方法の下準備として、正規分布する量の母集団があって、その N 個のデータから、母集団の平均値（**母平均**）μと分散（**母分散**）σ^2を推定することを考える。

得られたデータの算術平均（**標本平均**）$\bar{x}$ は、母平均μの**不偏推定量**（＝標本平均の期待値が母平均と等しい）である。しかし、標本平均は試行毎にばらつくので、母平均の値は一回の試行における標本平均を中心としてある範囲の中に存在すると言うことになる。確率的な現象を扱っているので、有限個のサンプル値からその範囲を 100%確実の意味で示すことはできない。そこで区間から外れる確率をαとし、値が大きい方に外れる確率と小さい方に外れる確率をそれぞれα/2 として、この間に入る確率 1-αの区間を定める。このような推定を**区間推定**と言い、そのようにして定められた区間を**信頼区間**と呼ぶ。当然ながら、αを小さくすると区間が広がり、区間を狭くすると外れる確率が大きくなる。通常、α=0.01 または 0.05 として、99%あるいは 95%の確率でその信頼区間内に母平均が含まれる設定が採られる（95%がより多く採用されている）。この範囲に母平均が含まれるのは一定の合理性があると考える、あるいは、この範囲から外れたら母平均の候補として棄却される、と言うような使い方になる。

この節での対象は、母平均、母分散共に未知なものとするが、解析の第一歩として、分散σ^2が既知の場合を考える。この場合、標本平均 $\bar{x}$ は、母平均μを中心に、分散σ^2/N の正規分布をする。100(1-α) %信頼区間を考えると、正規分布の累積確率の両端α/2 に囲まれる部分であり、95%信頼区間では、

$$\bar{x} - 1.96\sigma/\sqrt{N} \le \mu \le \bar{x} + 1.96\sigma/\sqrt{N} \quad (\alpha = 0.05) \tag{10.4}$$

と推定される。

次に、本来の目的である母分散σ^2 が未知の場合について議論を進める。母分散が未知の場合は、次式で与えられる不偏推定量（**不偏分散**：標本値から処理した量が母集団の分散に等しくなる量）$\hat{V}$ を(4.4)式のσ^2 の代わりに用いる。（注：分散の不偏推定の意味については、第 9 章 8 ページのティータイム欄参照）

$$\hat{V} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{10.5}$$

しかし、単に置き換えただけではだめである。なぜなら、σ^2(=V)は固定値であるが、$\hat{V}$ は確率分布する量であるからである。

$\bar{x}$ を正規化した変数 $\bar{x}_0$ は

$$\bar{x}_0 = \frac{\bar{x} - \mu}{\sqrt{V/N}} \tag{10.6}$$

となり、標準正規分布である。

一方、(10.6)式の V を$\hat{V}$ に置き換えた変数 t は次式となる。

$$t = \frac{\bar{x} - \mu}{\sqrt{\hat{V} / N}} \tag{10.7}$$

$\hat{V}$ と次式で関係付けられる以下の量 U

$$U \equiv \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{\sigma^2} = \frac{(N-1)\hat{V}}{\sigma^2} \tag{10.8}$$

は自由度 N-1 のカイ二乗分布である。(10.7)式の $\bar{x} - \mu$ と $\hat{V}$ を x_0 と U を用いて表すと、同式は

$$t = \frac{\bar{x} - \mu}{\sqrt{\hat{V} / N}} = \frac{\sigma x_0}{\sqrt{N\sigma^2 U / ((N-1)N)}} = \frac{x_0}{\sqrt{U / (N-1)}} \tag{10.9}$$

となる。これは、付録の式(10.31)の形であり、t は自由度 N-1 の t 分布

$$f(t;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{z^2}{n}\right)^{-(n+1)/2} \qquad (-\infty < t < \infty;\ n = N-1) \tag{10.10}$$

である。

ゆえに、信頼区間を決めるパラメータ αに対しては、式(10.33)で用いた表記 $t_{\alpha/2}$ の値を $t(n;\alpha)$と置いてこれを用いると、

$$\bar{x} - t(N-1;\alpha)\sqrt{\frac{\tilde{V}}{N}} \le \mu \le \bar{x} + t(N-1;\alpha)\sqrt{\frac{\tilde{V}}{N}} \tag{10.11}$$

と定められる。95%信頼区間（α=0.05）については、$t(n;0.05)$の値を付録の表 10.1 に挙げている。

10.3　回帰直線と信頼区間

（１）最小二乗法による回帰直線を求める

まず、回帰直線の決定法を述べ、その後に区間推定の方法を述べる。回帰直線の決定には最小二乗法を用いる。最小二乗法では図 10.3 に示す残差 $\varepsilon_i = y_i - \hat{y}_i$ の２乗和

$$S_{\varepsilon\varepsilon} = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i)^2 \tag{10.12}$$

が最小になるよう $\hat{a}, \hat{b}$ を定める。ここでの残差ε_i は回帰直線と観測値 y_i とのずれであり、(10.2)式で与えた誤差要因としての e_i とは違うものであることに注意してほしい。回帰直線の決定には、観測値から得られる次の１次処理データを用いる。

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i\ ,\ \ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{10.13a, b}$$

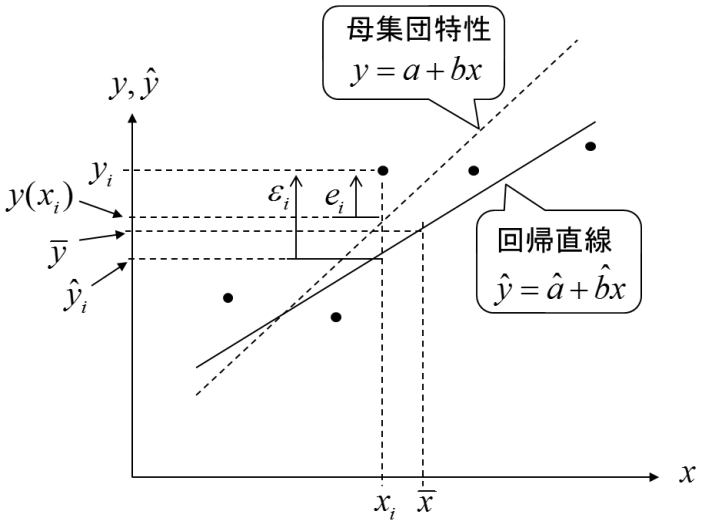


図 10.3　データの母集団特性と回帰直線（回帰直線は重心 $(\bar{x}, \bar{y})$ を通る）

$$\bar{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2, \quad \bar{y^2} = \frac{1}{N}\sum_{i=1}^{N} y_i^2 \tag{10.14a, b}$$

$$S_{xx} = \sum_{i=1}^{N}(x_i - \bar{x})^2 = N(\bar{x^2} - \bar{x}^2), \quad S_{yy} = \sum_{i=1}^{N}(y_i - \bar{y})^2 = N(\bar{y^2} - \bar{y}^2) \tag{10.15a, b}$$

$$S_{xy} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = N(\bar{xy} - \bar{x}\,\bar{y}) \tag{10.15c}$$

以下、最小二乗法で $\hat{a}, \hat{b}$ を定める。$S_{\varepsilon\varepsilon}$ の $\hat{a}, \hat{b}$ に対する最小値を決めるために $S_{\varepsilon\varepsilon}$ を $\hat{a}, \hat{b}$ でそれぞれ微分して、その値を 0 と置く。

$$\frac{\partial S_{\varepsilon\varepsilon}}{\partial \hat{a}} = \sum_{i=1}^{N} 2\varepsilon_i \frac{\partial \varepsilon_i}{\partial \hat{a}} = -2\sum_{i=1}^{N} \varepsilon_i = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i) = 0 \tag{10.16a}$$

$$\frac{\partial S_{\varepsilon\varepsilon}}{\partial \hat{b}} = \sum_{i=1}^{N} 2\varepsilon_i \frac{\partial \varepsilon_i}{\partial \hat{b}} = -2\sum_{i=1}^{N} \varepsilon_i x_i = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i)x_i = 0 \tag{10.16b}$$

上式は、以下の連立方程式に整理される。

$$\hat{a} + \hat{b}\bar{x} = \bar{y} \tag{10.17a}$$

$$N\hat{a}\bar{x} + (S_{xx} + N\bar{x^2})\hat{b} = S_{xy} + N\bar{x}\,\bar{y} \tag{10.17b}$$

これより、$\hat{a}, \hat{b}$ は以下のように定まる。

$$\hat{a} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \tag{10.18a}$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \tag{10.18b}$$

このように定めた$\hat{a},\hat{b}$は、二乗誤差を最小にすると共に、(10.16)式より、誤差の平均値もεx_iの平均値も 0 にしている。さらに、この回帰直線は、(10.17a)式より重心$(\bar{x},\bar{y})$を通ることも分かる。

係数$\hat{b}$は回帰直線の傾きを与えるが、相関係数ρとは

$$\hat{b} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\sqrt{\frac{S_{yy}}{S_{xx}}} = \sqrt{\frac{S_{yy}}{S_{xx}}}\rho \tag{10.19}$$

で関係付けられ、相関係数に比例する傾きとなることが分かる。

（２）信頼区間を求める

母集団特性推測の信頼区間を求める。このためには、回帰推定値$\hat{a},\hat{b}$ のそれぞれの分散を求めることから始める。先の議論と同様に、第１ステップとしてその分散が既知であるとする。

回帰係数$\hat{b}$は

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})\frac{1}{S_{xx}} = \sum_{i=1}^{N}\frac{x_i - \bar{x}}{S_{xx}}(y_i - \bar{y}) \tag{10.20}$$

であるので、$y_i - \bar{y}$の分散が既知（$=\sigma^2$）であるとすると、独立な変数の分散の加法性から、$\hat{b}$の分散$V_{\hat{b}}$は以下になる。

$$V_{\hat{b}} = \sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2\sigma^2 = \frac{\sigma^2}{S_{xx}} \tag{10.21}$$

$\hat{a}$の分散$V_{\hat{a}}$も同様に

$$V_{\hat{a}} = V_{\bar{y}} + \bar{x}^2 V_{\hat{b}} = \left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2 \tag{10.22}$$

となる。$\hat{a}$も$\hat{b}$も上記の分散を持つ正規分布をすると言うことになる。

しかし、分散σ^2は未知であって使うことができないため、10.2 節で述べたと同じように、次のステップでは実現値から得られる残差の不偏分散 $\hat{V}_\varepsilon$で置き換える。この議論をするためには、

カイ二乗分布における自由度の理解が重要になる。

三つの量：平均値との差の二乗和 S_{yy}, 回帰二乗和 $S_{\hat{y}\hat{y}}$, 残差二乗和 $S_{\varepsilon\varepsilon}$に着目する。

$$S_{yy}=\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}y_i^2-N\bar{y}^2 \tag{10.23a}$$

$$S_{\hat{y}\hat{y}}=\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{N}\hat{y}_i^2-N\bar{y}^2 \tag{10.23b}$$

$$S_{\varepsilon\varepsilon}=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\sum_{i=1}^{N}y_i^2+\sum_{i=1}^{N}\hat{y}_i^2-2\sum_{i=1}^{N}y_i\hat{y}_i \tag{10.23c}$$

残差自乗和 $S_{\varepsilon\varepsilon}$の式(10.23c)右辺の第３項は

$$\sum_{i=1}^{N}y_i\hat{y}_i=\sum_{i=1}^{N}(\hat{y}_i+\varepsilon_i)\hat{y}_i=\sum_{i=1}^{N}\hat{y}_i^2+\sum_{i=1}^{N}\varepsilon_i(\hat{a}+\hat{b}x_i)=\sum_{i=1}^{N}\hat{y}_i^2 \tag{10.24}$$

となる。最終辺は、前節で述べた最小二乗法の帰結である(10.16)式より、ε_i も$\varepsilon_i x_i$ もその総和が 0 となる性質を使っている。結局、

$$S_{\varepsilon\varepsilon}=\sum_{i=1}^{N}y_i^2-\sum_{i=1}^{N}\hat{y}_i^2 \tag{10.25}$$

となり、三つの量は、

$$S_{yy}=S_{\hat{y}\hat{y}}+S_{\varepsilon\varepsilon} \tag{10.26}$$

と関係付けられる。S_{yy} の自由度（S_{yy} は大きさが正規化されていないのでガンマ分布であるが、正規化するとカイ二乗分布になり、その意味での自由度）は N-1、$S_{\hat{y}\hat{y}}$ は自由度 1 であるので、$S_{\varepsilon\varepsilon}$は残りの自由度である N-2 (=N-1-1)である。

これによって、式(10.21), (10.22)の分散σ^2 を不偏分散$\hat{V}_\varepsilon$ で置き換える準備ができたことになる。その不偏分散は、上記より次式である。

$$\hat{V}_\varepsilon=S_{\varepsilon\varepsilon}/(N-2) \tag{10.27}$$

いよいよ、最終段階である区間推定に入る。回帰直線 $\hat{y}=\hat{a}+\hat{b}x$ を少し書き換えて

$$\hat{y}=\hat{a}+\hat{b}x=\bar{y}+\hat{b}(x-\bar{x}) \tag{10.28}$$

として、$\hat{y}$ の分散$V_{\hat{b}}$ は

$$V_{\hat{y}} = V_{\bar{y}} + (x-\bar{x})^2 V_{\hat{b}} = \left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} \right) \sigma^2 \tag{10.29}$$

となる。

このσ^2を自由度 N-2 の不偏分散$\hat{V}_\varepsilon$に置き換えると、信頼度 1-αの信頼区間（上限値と下限値）は、

$$\hat{a} + \hat{b}x \pm t(N-2,\alpha) \sqrt{\left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} \right) \hat{V}_\varepsilon} \tag{10.30}$$

となる。この式が、本章が主題とした区間推定を与える式になる。

繰り返し述べてきたように、(10.30)式は、真の特性がこの範囲に存在する確率が高い部分を示す幅であって、個々の観測値 y_i が分布する幅（ばらつき幅）ではない。

回帰分析の具体例を示す。元の特性は、a =3, b =1, σ=1 で、x は 1～５で等間隔に N =5, 10, 20, 100 としている（N=10 は図 10.1）。図 10.4 はこの分析結果を示す。図には、これまで述べてきた多くの特徴が表れており、それを感じ取ってほしい。

10.4　さらに理解を深めるために

（１）モデル選択の重要性

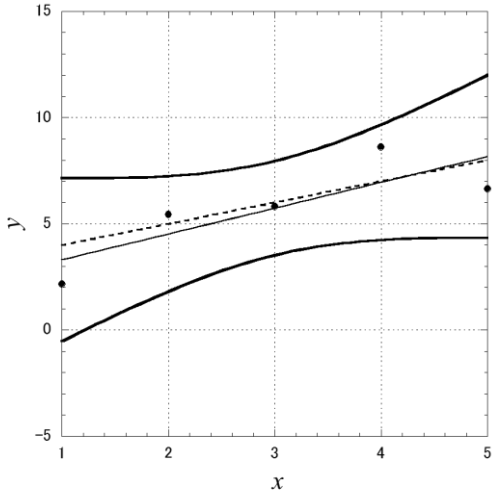
本節では、直線回帰の方法について詳しく述べたが、データを見てどのようなモデルを採用するかがスタート時点で、大事なことになる。選ぶモデルを間違うと、苦労して導いた結論が間違ったものになってしまう。そのためには、散布図を見て、特徴を見誤らないことである。その心構えとして、**アンスコムの例**（統計学者の Frank Anscombe が 1973 年に紹介した例）が有名である。アンスコムは、平均・標準偏差・相関係数が等しく、かつ、回帰直線も同じになる４つの異なるデータセット例（図 10.5）を示し、散布図をよく見てその傾向を確認することの重要性を指摘している。そして気をつけなくてはいけない点として、以下の３点を挙げている。

1) 曲線関係のあるデータに直線近似を当てはめる（採用モデルが間違っている）。（同図(b)）

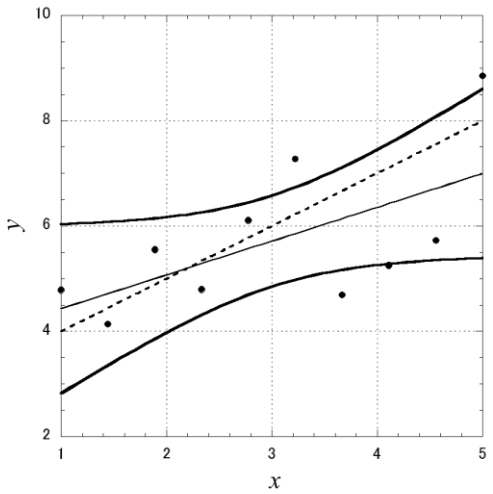
2) 直線状に並ぶデータの一点だけが異なっていて、カーブがその影響を受ける。この一点は何らかの理由によるはずれ値なので、データそのものを吟味する必要がある。（同図(c)）

3) 回帰直線が、離れている一点（はずれ値ではない）の影響を強く受けていて、統計的性質が正しく反映されていない可能性。（同図(d)）

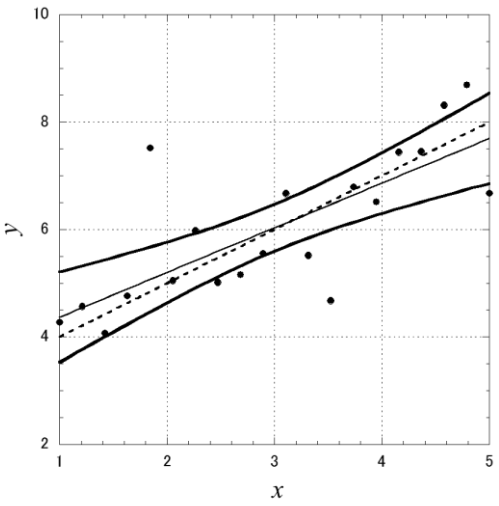
繰り返しになるが、アンスコムの例は、回帰分析を行う際には、データの統計値だけで比較するのではなく、散布図を見て、データの傾向や外れ値の有無など、解析のスタート時点において、慎重な確認が必要であることを主張している。



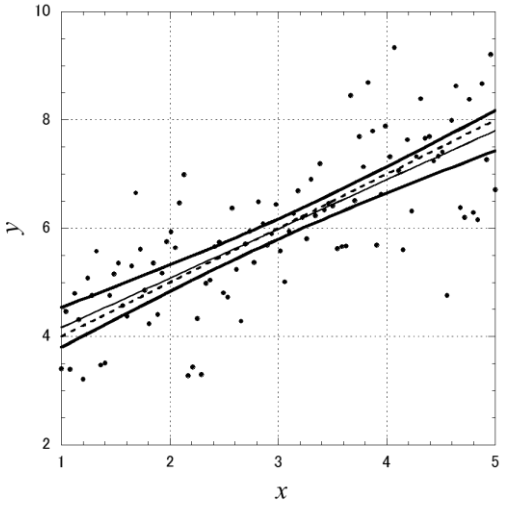
(a) N=5　（他と縦軸スケールが違う）



(b) N=10



(c) N=20



(d) N=100

図 10.4 回帰直線と 95%信頼区間（データ： a=3, b=1, σ=1; グラフ：点線：母集団特性、実線：回帰直線、太実線曲線：母集団特性の 95%信頼区間）

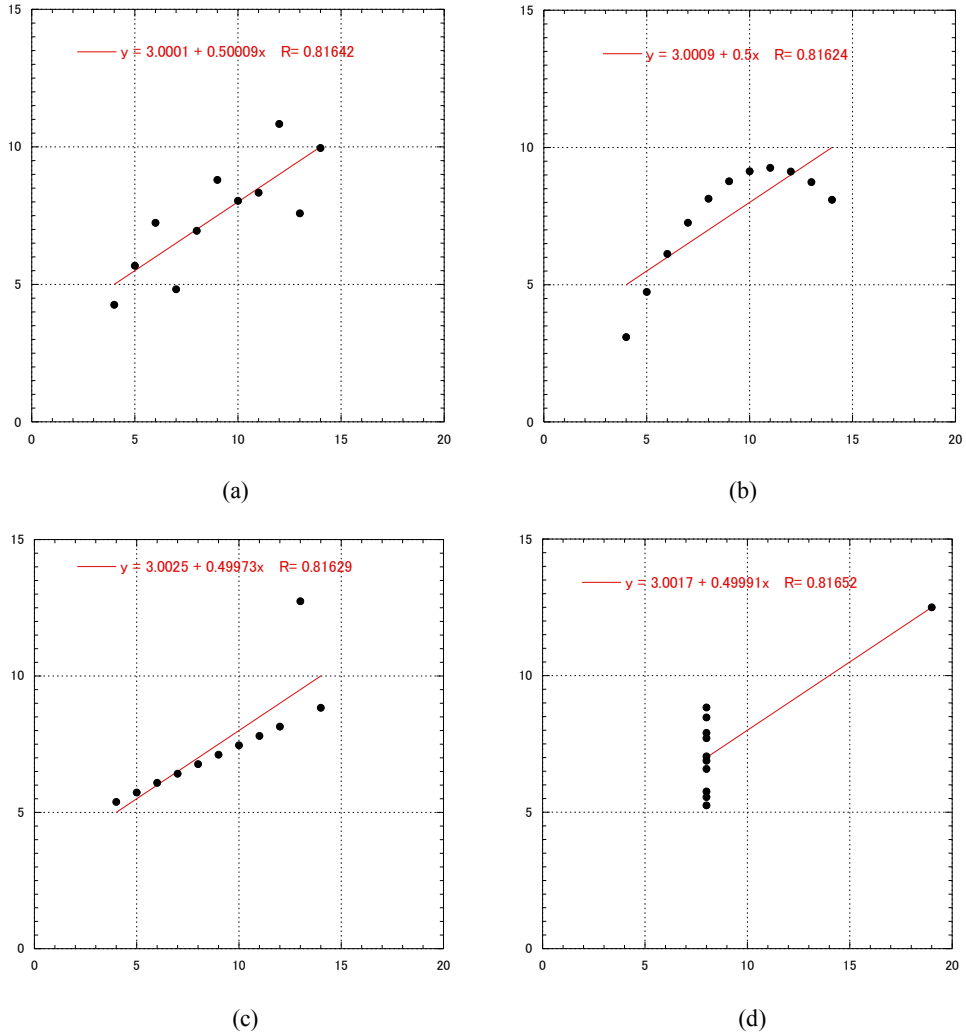


図 10.5　アンスコムの例（平均・標準偏差・相関係数が等しく、かつ、回帰直線も同じになる４つの例。散布図で確認すると違いが歴然。(a)が適切な適用例）

（２）x と y の関係

線形回帰を含む回帰分析では、x を原因とする量、y をその結果として現れる量とし、x には誤差が無いとして扱った。そのため、誤差は y 軸方向に発生し、それゆえ、誤差 e もその方向に採っている。そうすると、図 10.2 で示したようなケースでは、x と y が同じ確率分布をしているとしても、回帰直線は 45°の直線とならず、相関係数に応じて傾きが緩やかになる。

x と y とが、明確な原因・結果の関係（主従関係）にあれば、本節で述べた解析手法（回帰分析）でよいが、x も y も別の原因に支配されていて、単なる関係が有ると言う場合には、x も y も誤差を含むことになる。例えば、x も y も誤差の分散が同じ程度にあれば、誤差は点と回帰直線との距離で見る（すなわち、図 10.2 の例では点線で見る）よう変更しなければならない。このように、回帰分析では、横軸(x)、縦軸(y)の関係性の吟味が大事である。

観測値 y から原因 x を推定することは逆問題と言われる。回帰直線（あるいは曲線）とその信

頼区間で現された関係図があったとき、上記の変数の性質を理解して、逆問題をどのように推定するかは、各自で考えてほしい。

（３）補足

本章では、回帰分析の基本中の基本である２変量の直線回帰（単回帰モデル）とその信頼区間の定め方について、自己完結的に詳しくまとめた。次のステップとしては、２変量の一般的な関数（例えば多項式や指数あるいは対数など）への回帰分析、さらには、多変量での解析（多変量解析法）へと進んでゆくことになるであろう。特に後者では、様々な分野に溢れるビッグデータが来るべき AI 時代を支えており、データを見る目（＝その分析技術）を養っておくことは非常に重要である。

最近の市販の数値処理ソフト（エクセル、カレイダグラフなど）では、回帰分析機能が具備されていて、自分で計算しなくてもカーブを出してくれるが、その真偽（妥当性）や結果の物理的意味を見極めるためにも、基礎から原理を学んでおいてほしい。

10.5　信頼区間推定の電波気象統計への応用

統計的にものを見ることの大切さとして、信頼区間推定を取り上げた。少数のばらつきあるデータから、間違った答えを導かないよう判断の基準を与えてくれるものである。その一つの応用例として、日本に降る雨の一日あたり、あるいは、１時間あたりの量（日降水量・１時間降水量）の年間最大値の統計で見てみたい。降雨は高い周波数の電波に深刻な減衰をもたらすため無線通信には大敵であり、無関心ではいられない現象である。このような通信に影響を与える気象現象は**電波気象**と呼ばれ、通信分野での研究対象になっている。

日本各地（約 1,300 箇所）の気象データが気象庁のホームページから公開されている[3]。降水量に関しては、一日、１時間、１０分間に降る水の量、すなわち、日降水量、時間降水量、10 分間降水量について、観測開始から今日までのおよそ 100 年間のデータベースがある（期間は観測点やデータの種類によって異なる）。年間の最大値を与える降水現象は、例外なく降雨であるので、以下、降水量は降雨量と読み替える。

近年の気温の増加について、地球全体では 0.73℃/100 年の、日本（都市化による影響が比較的小さい 15 地点）では 1.21℃/100 年の増加が報告されている[4]。大都市を代表する東京（千代田区にある東京管区気象台エリア）での年間平均気温の上昇率は 3.2℃/100 年のペースで、日本各地の中で一番大きい。東京は都市化率が日本の都市の中で最大（92.9%）であり、地球温暖化に加え、都市構造の変化によるヒートアイランド現象も加わっているためと説明されている。図 10.6 は、100 年間(1920-2019)での、東京の毎年の平均気温を示している。区間推定によって、回帰直線の傾向が信頼できるものであることを示している。このように、ここ 100 年のスケールで見ると、確実に地球温暖化は進んでいる。では、降水量データで見るとどうであろうか？

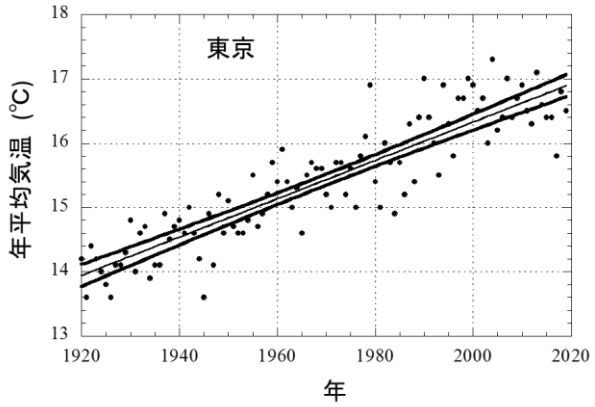


図 10.6　東京の年平均気温の推移（細線：回帰直線、太線：95%信頼区間）
（地球温暖化＋ヒートアイランド現象）

気温上昇率が我が国で一番高い東京について、１時間降水量の年間最大値を図 10.7 に示す。回帰直線を見ると僅かな上昇が認められるが、信頼区間推定では温暖化の影響が現われているとは言えないことになる。著者が調べた限りにおいて、実は、東京・大阪の大都市ばかりでなく、我が国の広い範囲に亘って、降水量（日・１時間）の年間最大値の特性に関しては同様であった[5]。これは、強雨の特性において、年毎のばらつきが大きいため、変化傾向が有意と判断できないという意味であり、100 年の長期間といえども、まだ、データが足りないのである。

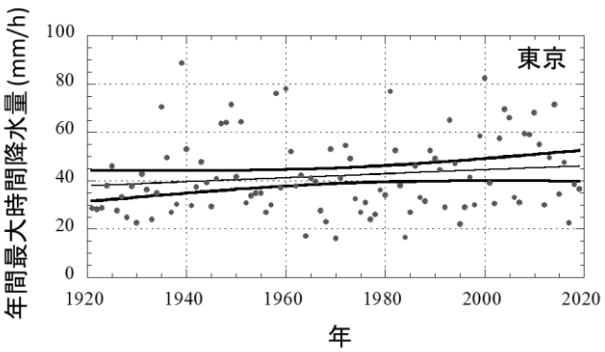


図 10.7　東京における時間降水量の年間最大値の年変化（年毎のばらつきが大きく、信頼区間ははっきりした傾向を論じることはできない幅を有している）

そこで、データ数を増やして傾向を浮かび上がらせたいため、気象庁が公開している気象観測データから、日本全国 45 地点・80 年間（1943～2023 年）の１時間降水量の年間最大値（データ点数 3600）を抽出した。そして、それぞれの地点ごとに 80 年での平均値を求め、それで割った正規化１期間降水量（平均値=1）を算出した。年変化傾向を見たい目的であるので、このように正規化し、データ数を増やしたのである。このデータを用いて１時間降水量の年間最大値の経年変化特性を求めたのが図 10.8（左側：全体、右側：縦軸拡大）である[6]。

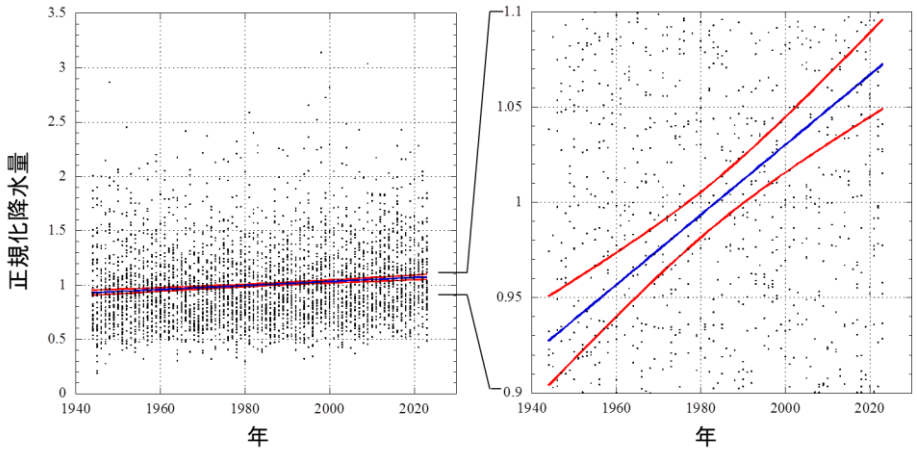


図 10.8　日本全国における正規化１時間降水量の年間最大値の分布と最尤推定（回帰直線）及び区間推定

同図の右側グラフより、データが増えたことによって信頼区間幅が狭まり、増加傾向が有意と判断されるのである。最尤推定（回帰直線）では、80 年間で約 1.16 倍（16%増）を示しているが、区間推定も 1.10（上限の左端の値と下限の右端の値の比）～1.21（下限の左端の値と上限の右端の値の比）となって、回帰直線の増加傾向をサポートしている。

ここでは、区間推定を利用して、降雨現象に対する長期的な変化傾向の一端を示したが、気象庁では、近年の強の降り方が変わってきている根拠として、1 年間に一定強度以上になる強雨の発生回数が増えてきているというデータを示している[7]。1976 年より、日本全国約 1300 地点において定常観測がスタートした地域気象観測システム（AMeDAS）のデータを用い、1 時間雨量が 50mm, 80mm, 100mm 以上になる発生回数の年変化特性を調べたものである。図 10.9 は、このうちの 50mm を超える強雨の発生回数に 95%信頼区間を付して示している。

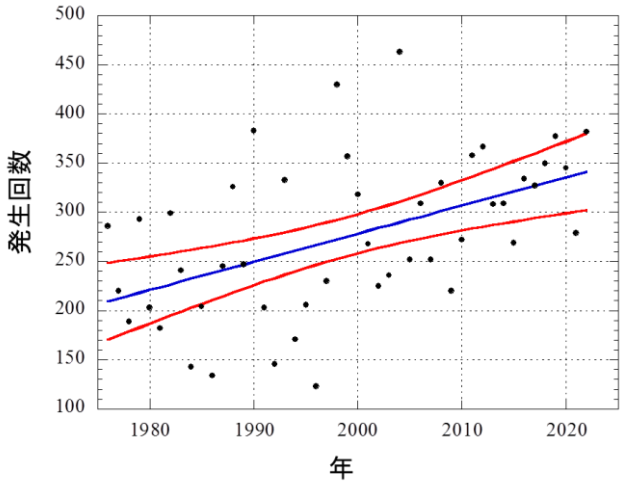


図 10.9　全国約 1300 地点当たりの 1 時間降水量 50mm 以上の大雨の年間発生回数
（回数の数値データは[7]より）

同図より信頼区間推定（信頼水準 95% : 赤線）でも増加は有意であることを示している（資料[7]では、しきい値 50mm, 80mm について信頼水準 99%で有意と記されている）。注意事項・補足として「これらの変化には地球温暖化が影響している可能性がある」と控えめに記されている。

このように、区間推定も利用の仕方を工夫すると、見えにくい傾向を自信を持って結論することができ、有力な判断手段であることが理解できると思う。

付録　t 分布

区間推定など統計や検定に用いられる t 分布を、確率分布を説明した本書第一部に入れていなかったので、ここで概要を説明する。

変数 x が標準正規分布 $N(0,1)$に従い、y が自由度 n のカイ二乗分布に従う独立な確率変数とする。このとき、

$$t = \frac{x}{\sqrt{y/n}} \tag{10.31}$$

の確率密度関数は次式となる（ベータ関数を用いた別の表現式もあるが同じもの）。

$$f(t;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \qquad (-\infty < t < \infty) \tag{10.32}$$

この分布を自由度 n の t 分布と呼ぶ。

この説明では n は自然数になるが、実数に拡張しても(10.32)式は確率密度関数の条件を満たしている。図 10.10 に n=1, 2, 5, 20 および標準正規分布の例を示している。式からも分かるように、t=0 を中心に正負対称の形である。n=1 では裾の広がりが大きいが、n の増加と共にその広がりが狭くなって正規分布に収斂することが分かる。（なお、n=1 はコーシー分布呼ばれる）

本文で述べた区間推定では、両側α/2 ずつの累積確率を除いた部分の確率を与える次式

$$\int_{-t_{\alpha/2}}^{t_{\alpha/2}} f(t;n)\,dt = 1-\alpha \quad (0<\alpha<1) \tag{10.33}$$

においては、確率αを与えて $t_{\alpha/2}$ を求める必要が出てくる。これは解析的に解くことができないため統計の本の付録等に載っている表から読み取る必要がある。表 10.1 に 95%信頼区間を求めるのに必要なα=0.05 のときの n=1～30 に対する $t_{\alpha/2}$ の値をまとめている。これ以上に n が大きくなると、標準正規分布の値である 1.960 に漸近してくる。n=30 ($t_{\alpha/2}$=2.042)以上では、標準正規分布で代用しても問題となる誤差にはならない。ちなみに 95%水準の漸近値 1.960（n=∞）に対応する数値として、90%水準では 1.645、99%水準では 2.576 である。

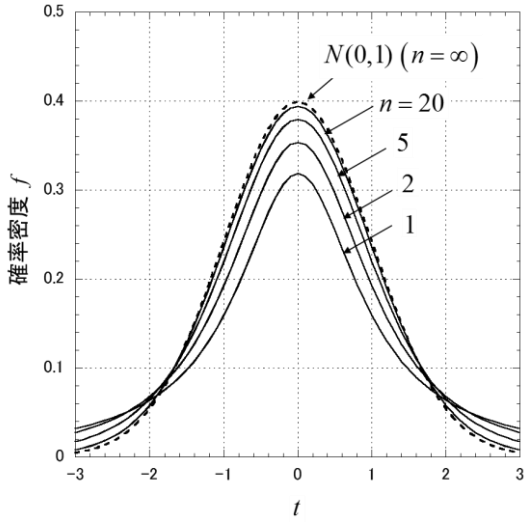


図 10.10　t 分布の確率密度関数

表 10.1　t 分布のα=0.05 のときの $t_{\alpha/2}$ の値

n =1	2	3	4	5	6	7	8	9	10	
12.706	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262	2.228	
11	12	13	14	15	16	17	18	19	20	
2.201	2.179	2.160	2.145	2.131	2.120	2.110	2.101	2.093	2.086	
21	22	23	24	25	26	27	28	29	30	∞
2.080	2.074	2.069	2.064	2.060	2.056	2.052	2.048	2.045	2.042	1.960

参考文献

[1] 唐沢好男, 無線通信物理層技術へのアプローチ, コロナ社, 2021.

[2] 芳賀敏郎, 医薬品開発のための統計解析：第 1 部 基礎 改訂版, サイエンティスト社, 2011.

[3] 気象庁, 過去の気象データ検索, 気象庁ホームページ, https://www.data.jma.go.jp/obd/stats/etrn/index.php

[4] 気象庁, “地球温暖化予測情報 第 9 巻,” 2017.03. https://www.data.jma.go.jp/cpdinfo/GWP/Vol9/pdf/all.pdf

[5] 唐沢好男, “日本の降雨量極値データを統計的に見る（改訂版）, 技術レポート（私報）YK-021-rev, 2019. http://www.radio3.ee.uec.ac.jp/ronbun/YK-021_Rainfall_Statistics.pdf

[6] 唐沢好男, “日本の降雨極値データに見る長期変化傾向, 技術レポート（私報）YK-082, 2024. http://www.radio3.ee.uec.ac.jp/ronbun/YK-082_Rainfall_Statistics_2.pdf

[7] 気象庁, 大雨や猛暑日（極端現象）のこれまでの変化, 気象庁ホームページ. https://www.data.jma.go.jp/cpdinfo/extreme/extreme_p.html

目次のページはこちら