Technical Report YK-019
Jan. 8, 2019
Y. Karasawa

回帰分析と信頼区間

～ばらつきの大きい少数データから誤った推論をしないための～

唐沢好男

電波伝搬に関連するデータをグラフにすると、大きなばらつきを持つことがよくある。その傾向を見たいとき、年期を積んだその道のプロは、勘によってエイヤッと線を引き、なるほどと思える結果を見せてくれる。しかし、そういう職人芸が常に良いとは限らず、誰でも間違いなくできる統一的な手法が求められる。その有力な数学手段が最小二乗法であり、横軸（原因）と縦軸（結果）の関係性を示すカーブ（回帰曲線と言われる）を合理的に定めてくれる。そのようにして得られるカーブであっても、データ数が少なかったり、ばらつきが大きかったりすると、データの組毎に異なったものとなってしまう。この回帰曲線の信頼性を定量的に示すのが信頼区間である。得られたデータから間違った推論をしないために生まれた概念である。本レポートでは、データの統計的処理において最も基本となる直線回帰（単回帰モデル）にポイントを絞り、信頼区間の考え方や求め方を基礎から丁寧に示す。

目次

Technical Report YK-019
Jan. 8, 2019
Y. Karasawa

１．回帰直線と信頼区間：その概要

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（注 1)。統計で言うところの母集団の真の特性である。

$$y(x) = a + bx \tag{1}$$

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。

$$y_i = a + bx_i + e_i \tag{2}$$

N 個のデータの処理によって得られる回帰直線

$$\hat{y}(x) = \hat{a} + \hat{b}x \tag{3}$$

は、母集団の特性、すなわち、(1)式の y を推定するものになる。この(1)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[2]、本レポートでもそれに従う（注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）

図 1 は本レポートでの要点を説明するための模式的な図である。

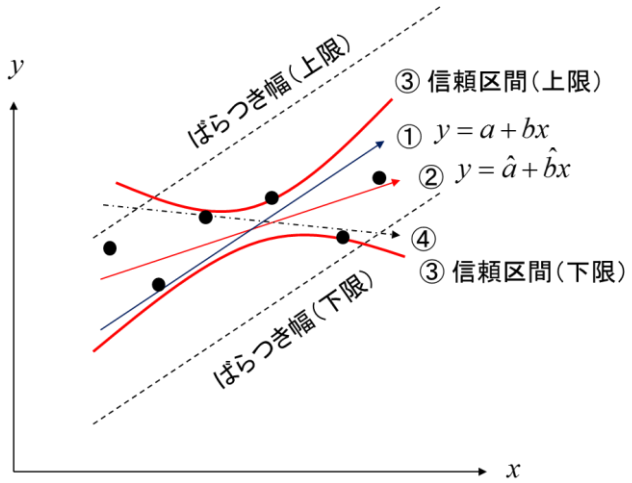


図１　ばらつきのある少数データとその回帰直線のイメージ図（データに基づく正確な図は、本レポートの最後に示す（図８)。図中のばらつき幅はデータそのもののばらつきの意味）

図１では、①の関係を持つ事象に対して、雑音が加わってばらつきのある６つのデータをプロットしている。最小二乗法によって求めた回帰直線が②である。②はあくまでも、得られたデータに対して誤差の最小化を行った直線なので、データが変わると回帰直線も変わってしまう。データ数 N を十分多くすることによって、②は①に漸近してゆくが、図１はまだその途中の状態である。②を見て、「y は x が大きくなると大きくなる傾向がある」と言ってよいのだろうか。これに答えを与えるのが、信頼区間の考え方である。信頼区間とは、予め定めた確率に対して、真の特性①がその範囲に入っていると推量できる区間を言う（注３）。③のカーブが信頼区間の上限と下限を与えている。信頼確率を大きくとれば信頼度は上がるが、信頼区間の幅が広くなってしまうので、どこかで適当な確率を定める必要がある。多くの場合、その確率の値は 95%に定められる。信頼区間内であれば、それに含まれる特性、例えば④の仮説（y は x の増加に対して減少する）も排除できないことになる。当然ながら、信頼区間の幅は、データ数が増えるとそれに応じて狭くなり、より精度のよい推定が可能になる。信頼区間とは、母集団の真の特性（(1)式で与えられる）が指定確率で存在する範囲を与えるものであって、図の点線に示すような母集団データのばらつき幅を意味するものではない。

本レポートでは、３章で母平均の統計的推定法を、４章で単回帰モデルに対する回帰分析の具体的な方法を述べる。２章ではその準備ための道具立てを述べる。

［注１］確率・統計の教科書等では、確率変数は大文字 X, Y で、その実現値は小文字 x, y で表記することが一般的であるが、ここではそれを区別せずに、共に小文字で与えている。（区別すると、表記に混乱が起きそうな事例もあるため）

［注２］明確な原因と結果の関係であれば、i)の仮定は合理的であるが、単に、x と y には（他の本質的な変数を介して）関連性があると言うような場合には、吟味が必要である。また、結果から原因を推定するような逆推定問題ではこの前提が逆になる。これらの議論については、第５章で触れる。

［注３］95%信頼区間が具体的に定まったとき、「母平均が 95%の確率でその範囲に存在する」と考えるのは、正確には正しくない。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（毎回変動する）に存在する」と考えた方がより正しい意味になる。

【「回帰」の由来】

本レポート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。基本的な用語の定義や説明、図の作り方などは、次節以降で行うため、ここでは、眺めるように見てほしい。

図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自乗法で求めると、点線よりは傾きが小さい赤線のように求められる。赤線と楕円の交点の楕円部分は傾きが垂直になっている。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、次章に進んでほしい。

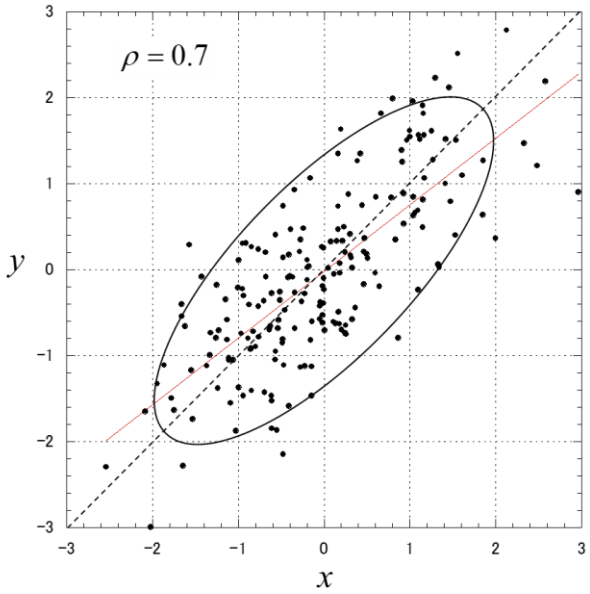


図２　相関を有する二つの正規乱数の散布図と回帰直線（赤線）

２．数学的準備

３章、４章で、信頼区間の求め方を述べる。本章はその準備のための基礎的な項目（解析のための道具立て）をまとめる。ここでは、項目の記述に留め、導出過程や証明を省いている。それらの詳細については確率・統計の教科書や専門書で学んでほしい。筆者も無線通信に現れる確率分布について、[5], [6]に、より詳しくまとめているので、併せて見てほしい。

２．１　平均と分散

期待値 μ

確率変数 x が、確率密度関数 $f(x)$に従うとき、x の期待値μは

$$\mu\{\equiv< x >\} = \int_{-\infty}^{\infty} xf(x)dx \tag{4}$$

で求められ、母集団の平均値である。（本稿では<・>を期待値を表す記号に用いる。$E(\cdot)$で表される場合も多い。）【注：(4)式の $f(x)$は連続関数を想定するが、離散値を含む場合でもデルタ関数を用いて表せば、連続・不連続を含んで通用する。また、この章と３章では確率変数として文字 x を代表として使うが、4 章で扱う基準変数ではなく、一般的な意味である。】

標本平均 $\overline{x}$

母集団から N 個を取り出しそれぞれを x_i（i=1, 2, …, N）とするとき、その算術平均値$\overline{x}$は

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

で求められ、標本平均と呼ばれる。標本平均の期待値は母平均と等しい。すなわち

$$\langle \overline{x} \rangle = \mu \tag{6}$$

このように期待値が母集団の統計値と一致することを不偏性が有ると言い、$\overline{x}$ は不偏推定量である。

分散 V, 標準偏差 σ

母集団の分散 V は次式で定義される。

$$V\{\equiv Var(x) \equiv< (x-\mu)^2 >\}$$
$$= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 \tag{7}$$

標準偏差σは分散 V の平方根である。

$$\sigma = \sqrt{V} \tag{8}$$

標本平均の分散

標本平均 $\bar{x}$ はサンプルされたデータに依存してばらつきがある。その分散は

$$Var(\bar{x})\left(\equiv\left\langle\left(\bar{x}-\mu\right)^2\right\rangle\right)=\frac{\sigma^2}{N} \tag{9}$$

となり、サンプル数 N の増加と共に 0 に近づく。

標本分散 v^2 と不偏分散 $\hat{V}$

サンプルされたデータの分散は標本分散と呼ばれ、

$$v^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2=\frac{1}{N}\sum_{i=1}^{N}x_i^2-\bar{x}^2 \tag{10}$$

である。この分散は、式中の標本平均 $\bar{x}$ がデータの分散を小さくする方向に揺れているため、母集団の分散 σ^2 を予測するものとしては過小評価になる。標本分散の期待値を求めると

$$\left\langle v^2\right\rangle=\frac{1}{N}\sum_{i=1}^{N}\left\langle x_i^2\right\rangle-\left\langle\bar{x}^2\right\rangle=\sigma^2+\mu^2-\frac{\sigma^2}{N}-\mu^2=\frac{N-1}{N}\sigma^2 \tag{11}$$

となり、母集団の分散 σ^2 より比率で $(N\text{-}1)/N$ だけ小さい値だったと言うことになる。このため、標本値から母集団の分散を推定するためには、

$$\hat{V}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2 \tag{12}$$

としなければいけない。このようにして求めた分散は不偏分散（あるいは不偏標本分散）と呼ばれ、母集団の x の分散 σ^2 に対する不偏推定量である。このため、母集団の分散推定に際して、少数データの処理では、標本分散より不偏分散が用いられる。

最小二乗法と標本平均の関係

サンプル値 x_i（i=1, 2, …, N）と定数値 m との差 e_{mi} の 2 乗値の和を求める。

$$S=\sum_{i=1}^{N}e_{mi}^2=\sum_{i=1}^{N}(x_i-m)^2 \tag{13}$$

S を最小とする m を求めたいために S を m で微分して 0 と置くと次式になる。

$$\frac{dS}{dm} = 2\sum_{i=1}^{N} e_{mi}\frac{de_{mi}}{dm} = -2\sum_{i=1}^{N} e_{mi} = -2\sum_{i=1}^{N}(x_i - m) = 0 \tag{14}$$

これより、S の極小値を与える m は

$$m = \sum_{i=1}^{N} x_i = \overline{x} \tag{15}$$

となる。母平均の不偏推定値である $\overline{x}$ はサンプルデータの最小二乗誤差となる数値でもある。回帰分析では**最小二乗法**が有力な手段になるが、その理由はここにある。

２．２　相関と共分散

物理量 x が（時間等により）変化するとき、別の量 y も変化する現象がある。このとき、x と y の変化の似通い具合（類似度）を表す指標として相関が用いられる。人間の身長と体重のように、身長が増加すれば、体重も増加する傾向が強いようなものに対しては相関が強く、「風が吹くと桶屋が儲かる」的な因果関係の薄いものに対しては、相関が弱い、と言うように用いる。

最初に、母集団の変数 x, y に対する確率論的な定義から入る。定常確率過程においては、x, y の平均値をμ_x, μ_y、標準偏差をσ_x,σ_y とするとき、相関係数ρは以下の式で与えられる。

$$\rho = \frac{\left\langle (x-\mu_x)(y-\mu_y)\right\rangle}{\sqrt{\left\langle (x-\mu_x)^2\right\rangle\left\langle (y-\mu_y)^2\right\rangle}} = \frac{\sigma_{xy}^2}{\sigma_x\sigma_y} \tag{16}$$

ここで、σ_{xy}^2 は共分散と呼ばれる量で、次式で定義される。

$$\sigma_{xy}^2 \equiv \left\langle (x-\mu_x)(y-\mu_y)\right\rangle = \left\langle xy\right\rangle - \mu_x\mu_y \tag{17}$$

相関係数は、x と y の変化が完全相似形のとき 1 （完全相関）、変動方向が正負反転した完全相似形のとき -1 （負の完全相関）、独立なとき 0（無相関）となり（注１）、この範囲（-1～1）に値を持つ。$|\rho|$が 1 を超えないことはシュワルツの不等式により保証されている。

次に、サンプル値に対する相関係数 r を求める。この場合、母集団の相関を推定するには、標本分散でなく不偏分散を使う必要がある。不偏共分散も式(12)と同様に N-1 で割ることになるので以下の式で表される。

$$r = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2\right)\left(\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y})^2\right)}} = \frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\overline{x})^2\sum_{i=1}^{N}(y_i-\overline{y})^2}} \tag{18}$$

右辺の式（すなわち、最終的な形）は、普通に教科書等で見る式であり、中辺で用いている不偏分散を意識しなくてもよかったわけであるが、手順を踏んで考えれば、こうなっていると言うことを気に留めておくとよいであろう。サンプルデータの相関係数 r は、データ毎に揺らぐが、その期待値はρになる。

【注 1 ：無相関（相関係数が 0）には二つのケースがある。お互いが無関係であれば、統計的にランダムであり、無相関の一つのケース。もう一つは、sin と cos の変動のように、関係性は有るが直交している（内積が 0 になる）ケース。本レポートのテーマでは、前者の意味で用いるが、「無相関=無関係」ではないことを意識しておくとよい】

２．３　中心極限定理

確率変数 $x_1, x_2, \ldots, x_n$ が互いに独立に同一の確率分布に従い、その平均をμ、分散をσ^2 とする。このとき、$x_1, x_2, \ldots, x_n$ の平均：$\overline{x} = (x_1+x_2+\ldots+x_n)/n$ の確率分布は、n が十分大きければ、平均μ、分散σ^2/n の正規分布となる（正規分布の関数形は次節）。元の分布の形を問わないことが味噌で、意味するところが深い。例外として、コーシー分布のように平均や分散が定まらない分布（いわゆる、裾の広がりが大きい分布）では成立しないが、それは特殊なものであって、自然界で普通に現れる分布についてはこの定理が成立する。その証明は確率論の道具である特性関数を用いて行われる。

互いに独立に同一の確率分布であることを前提とした説明を行ったが、異なる分布であってもそれぞれが十分多数であれば、それでも正規分布に収斂する堅牢性が中心極限定理にはある。

複雑な要因が多数加わって実現しているような物理量、すなわち、加法性の確率過程に従って生み出されるような量は、中心極限定理が働いて正規分布に、あるいはそれに近いものになる。多数の要因が掛け算によって生起する乗法性の確率過程では、積の対数をとったものが加算で現されるので、この現象による物理量は対数正規分布する[5]。

２．４　回帰分析に重要な確率分布

２．４．１　正規分布

正規分布の確率密度関数は以下の式で与えられる。ガウス分布とも呼ばれる。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \equiv N(\mu, \sigma^2) \tag{19}$$

パラメータμは平均値、σは標準偏差である。正規分布を $N(\mu,\sigma^2)$の形で簡易に表すことも多い。

中心極限定理で述べられているように、正規分布は加法性確率過程での基本中の基本に位置する分布である。その中でも、平均値 0、標準偏差 1 の正規分布 $N(0,1)$は標準形、ある

いは、標準正規分布と呼ばれ、特に重要である。任意の正規分布 $N(\mu,\sigma^2)$は、確率変数 x を変数変換

$$y=\frac{x-\mu}{\sigma} \tag{20}$$

することで、確率変数 y に対しては $N(0,1)$の標準正規分布になる。正規分布 $N(0,\sigma^2)$に対する n 次モーメント（あるいは、$N(\mu,\sigma^2)$に対する中心モーメント）は、n=1,2,3,4 に対して、それぞれ、$0, \sigma^2, 0, 3\sigma^4$となる。

累積分布関数 F は次式である。

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx=\frac{1}{2}\left\{1+\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right\} \tag{21}$$

ここで、erf(・)は誤差関数で、次式で定義される。

$$\mathrm{erf}(x)\equiv\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{22}$$

正規分布には、正規分布する二つの変数の和の分布は正規分布になると言う再生性がある。

図３は標準型正規分布 $N(0,1)$の確率密度関数と累積分布関数を示している。信頼区間評価では、分布の端に値が存在する確率（図に影をつけた部分のような）が重要になる。注目される値としては、片側確率（1-$F(x)$）では、x=1,2,3 (σ)に対しては、0.1587, 0.0228, 0.0013 になる。両側確率（2(1-$F(x)$）では、それらの２倍になる。特異な存在の代表として 3σが使われるが、だいたい 1/1000 のことである。

次節で述べる区間推定では、中心から確率 1-αの中に入る x の下限（$-x_{\alpha/2}$）と上限（$x_{\alpha/2}$）が求められる。すなわち、

$$\int_{-x_{\alpha/2}}^{x_{\alpha/2}} f(x)\,dx=1-\alpha \quad (0<\alpha<1) \tag{23}$$

において、確率αを与えて $x_{\alpha/2}$ を求める必要が出てくる。通常用いられる 95%信頼区間（α=0.05）の場合は $x_{\alpha/2}=1.96\sigma$、99%信頼区間（α=0.01）では 2.58σとなる。

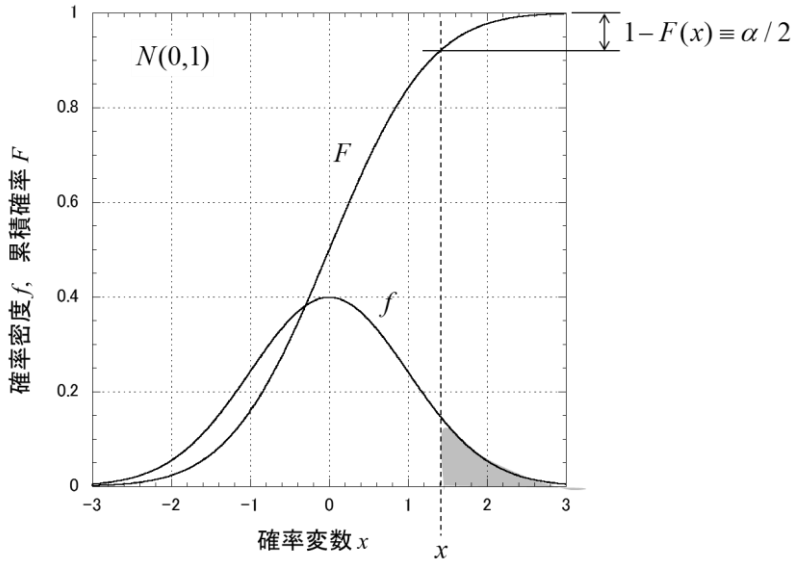


図３　標準正規分布の確率密度関数と累積分布

２．４．２　２次元正規分布

二つの正規分布系列 x, y があり、それぞれの平均値と標準偏差は、μ_x, μ_y およびσ_x,σ_y、相関係数ρで結ばれているとする。この２変数の結合確率密度関数 $f(x,y)$は次式で表され、２次元正規分布と呼ばれる。

$$f(x,y;\mu_x,\sigma_x,\mu_y,\sigma_y,\rho)$$
$$=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}\right] \quad (24)$$

お互いの変動が無相関の場合、ρ=0 なので、結合確率密度関数は、それぞれの確率密度関数（＝１次元正規分布）の積になる。

正規分布系列 $x(t)$、$y(t)$を相関係数ρとなるように生成したい場合には、独立な２つの標準正規分布系列 $u(t), v(t)$を用いて、次式により生成する。

$$\begin{pmatrix}x(t)\\y(t)\end{pmatrix}=\begin{pmatrix}\sigma_x & 0\\0 & \sigma_y\end{pmatrix}\begin{pmatrix}1 & 0\\\rho & \sqrt{1-\rho^2}\end{pmatrix}\begin{pmatrix}u(t)\\v(t)\end{pmatrix}+\begin{pmatrix}\mu_x\\\mu_y\end{pmatrix} \quad (25)$$

系列数が多い場合でも、その相互の相関係数を相関行列で与えることにより、設定した相関を有する複数ランダム系列を作ることができる。具体的には、文献[5]の 7.3.2 節を見てほしい。

２次元正規分布は横軸 x、縦軸 y として散布図で表すと関係をつかみやすい。両変数とも標準正規分布で、相関係数ρを 0.7 としたものは、図２で示している。同様に、ρ=0.9 と-0.5 としたものを図４に示す。図には、約 90%のデータが含まれる楕円を示している。この楕円

は、標本平均値$\bar{x},\bar{y}$、不偏標準偏差（不偏分散の平方根）$\hat{\sigma}_x,\hat{\sigma}_y$、データから得た相関係数$r$（$\rho$で設定している場合はその値）を用いて、以下の手順により求められる（図５）[2]。

1) 平均値（$\bar{x},\bar{y}$）を通る十字線を引く
2) $\bar{x}\pm 2\hat{\sigma}_x,\bar{y}\pm 2\hat{\sigma}_y$ の４辺形を描く
3) ４辺形の上に楕円の接点の位置：$\bar{x}\pm 2r\hat{\sigma}_x,\bar{y}\pm 2r\hat{\sigma}_y$ に×印をつける（４箇所）
4) 十字線上の$\bar{x}\pm 2\sqrt{1-r^2}\hat{\sigma}_x,\bar{y}\pm 2\sqrt{1-r^2}\hat{\sigma}_y$ に○印をつける（４箇所）
5) 上に求めた 8 つの点を通る楕円を描く（PowerPoint のような図形ソフトには、楕円のテンプレートがあるので、それがそのまま使える）

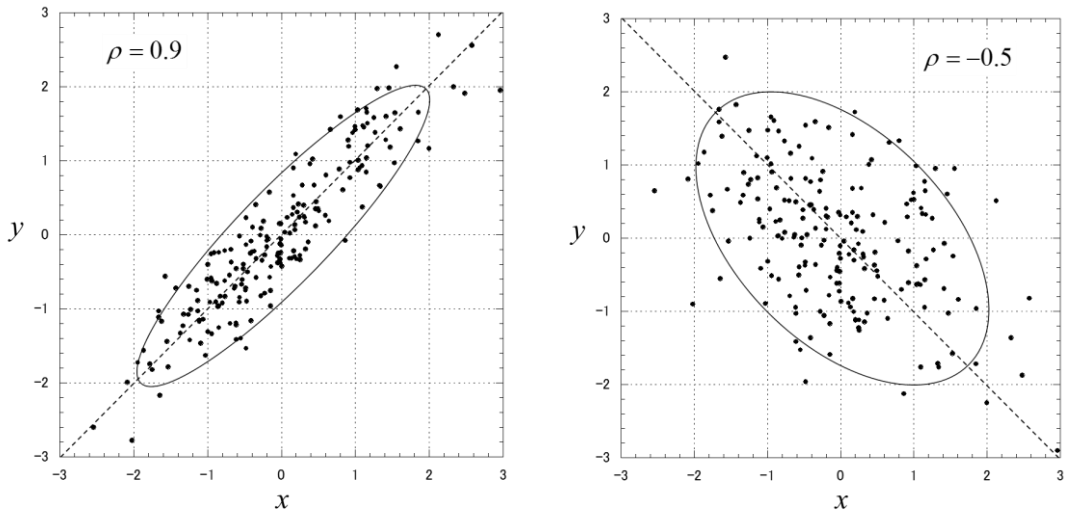


図４　相関係数をパラメータとした２次元正規分布の散布図

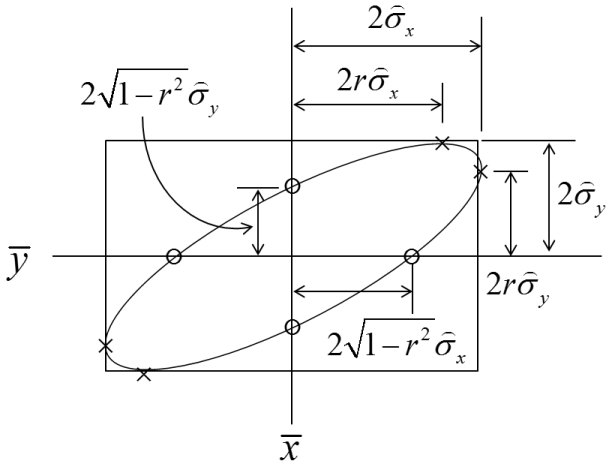


図５　約 90%のデータが含まれる楕円の求め方の説明図

２．４．３　χ^2 分布とガンマ分布

確率変数 $x_1, x_2, \ldots, x_n$ が各々標準正規分布 $N(0,1)$ に従い、かつ、それらが独立であるとき、それらの二乗和 $z=x_1^2+x_2^2+\ldots+x_n^2$ の分布は自由度 n の χ^2 分布（カイ二乗分布）と呼ばれる。その確率密度関数は次式である。

$$f(z;n)=\frac{1}{2\Gamma(n/2)}\left(\frac{z}{2}\right)^{n/2-1}\exp\left(-\frac{z}{2}\right) \qquad (z\geq 0)\ (\Gamma:\ \text{ガンマ関数}) \tag{26}$$

χ^2 分布には、自由度 n の χ^2 分布と自由度 m の χ^2 分布の和の分布は自由度 $n+m$ の χ^2 分布になるという再生性がある。

χ^2 分布は標準正規分布に対して与えられるものなので、スケールパラメータ β を導入し、かつ、$n/2$ を形状パラメータ ν に置き換えて表わすとガンマ分布となり、式(27)が得られる。

$$f(z)=\frac{1}{\Gamma(\nu)}\beta^{\nu}z^{\nu-1}\exp(-\beta z) \tag{27}$$

２．４．４　t 分布

変数 x が標準正規分布 $N(0,1)$ に従い、y が自由度 n の χ^2 分布に従う独立な確率変数とする。このとき、

$$t=\frac{x}{\sqrt{y/n}} \tag{28}$$

の確率密度関数は次式となる（注１）。

$$f(t;n)=\frac{1}{\sqrt{N}B\left(\frac{n}{2},\frac{1}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \qquad (-\infty<t<\infty) \tag{29}$$

ここで、$B(p,q)$ はベータ関数で、

$$B(p,q)=\int_0^1 x^{p-1}(1-x)^{q-1}dx \tag{30}$$

である。この分布を自由度 n の t 分布と呼ぶ。【この分布の発案者は William Sealy Gosset（ゴセット）。1908 年、論文を Student というペンネームで発表したため、スチューデントの t 分布とも呼ばれる。】

この説明では n は自然数になるが、実数に拡張しても(29)式は確率密度関数の条件を満たしている。図 6 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）

次節の区間推定では、両側$\alpha/2$ずつの累積確率を除いた部分の確率を与える次式

$$\int_{-t_{\alpha/2}}^{t_{\alpha/2}} f(t;n)\,dt = 1-\alpha \quad (0<\alpha<1) \tag{31}$$

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。これを表記上、

$$t_{\alpha/2} = t(n;\alpha) \tag{32}$$

と置く。表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。

（注１）t 分布の別表現

式(29)はガンマ関数(Γ)を用いて、以下のようにも表される[3]。（表現形式が違うだけなので、数値計算値は同じになる）

$$f(z;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{N\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{z^2}{n}\right)^{-(n+1)/2} \quad (-\infty<z<\infty) \tag{33}$$

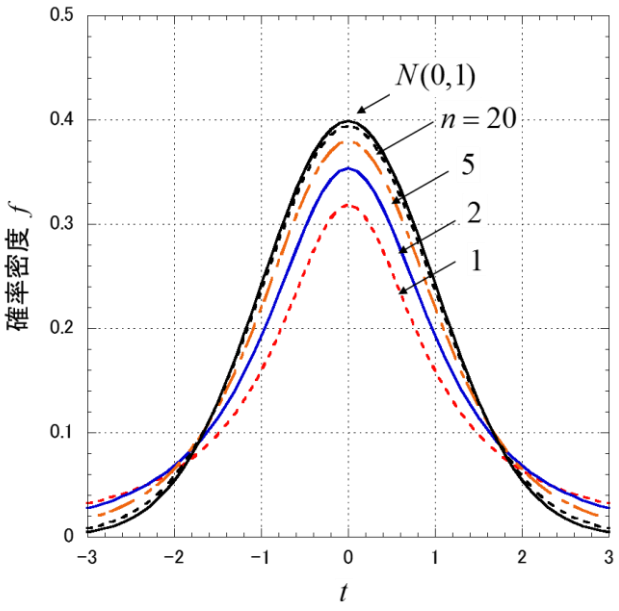


図６　t 分布

表１　t 分布のα=0.05 のときの $t_{\alpha/2}$ の値: $t(n;0.05)$

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

３．母平均の統計的推定

正規分布する量の母集団があって、そこから N 個のデータを取り出す。そのデータから、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。

2.1 節で述べたように、得られたデータの算術平均（標本平均）$\bar{x}$ は、母平均μの不偏推定量であることを述べた。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（注）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。

【注：１章の（注３）でも書いたように、95%信頼区間が１セットのデータによって $x_{min}\leq\mu\leq x_{max}$ と定まったとき、「母平均が 95%の確率でその範囲に存在する」と考えるのには、正確には正しくない。なぜなら、μが固定値であるのに対して、範囲を決める x_{min}, x_{max} が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、確率的に考えて、そのうちの 95 回は母平均がその都度の 95%信頼区間の範囲に存在する」と考えた方がより正しい意味になる。】

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、(9)式より、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、

$$\bar{x}-1.96\sigma/\sqrt{N}\leq\mu\leq\bar{x}+1.96\sigma/\sqrt{N}\quad(\alpha=0.05)\tag{34}$$

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、(12)式で与えられる不偏推定量（不偏分散）$\hat{V}$ を(34)式のσ^2 の代わりに用いる。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

$$\bar{x}_0 = \frac{\bar{x} - \mu}{\sqrt{V / N}} \tag{35}$$

となり、標準正規分布である。

一方、(35)式の V を$\hat{V}$ に置き換えた変数 t は次式となる。

$$t = \frac{\bar{x} - \mu}{\sqrt{\hat{V} / N}} \tag{36}$$

$\hat{V}$ と次式で関係付けられる以下の量 U

$$U \equiv \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{\sigma^2} = \frac{(N-1)\hat{V}}{\sigma^2} \tag{37}$$

は自由度 N-1 のχ^2 分布である。(36)式の $\bar{x} - \mu$ と $\hat{V}$ を $\bar{x}_0$ と U を用いて表すと、同式は

$$t = \frac{\bar{x} - \mu}{\sqrt{\hat{V} / N}} = \frac{\sigma x_0}{\sqrt{N\sigma^2 U / ((N-1)N)}} = \frac{x_0}{\sqrt{U / (N-1)}} \tag{38}$$

となる。これは式(28)の形であり、t は自由度 N-1 の t 分布をすることが分かる。

ゆえに、信頼区間を決めるパラメータαに対しては、式(32)で用いた表記$t_{\alpha/2} = t(n;\alpha)$ を用いて、

$$\bar{x} - t(N-1;\alpha)\sqrt{\frac{\tilde{V}}{N}} \le \mu \le \bar{x} + t(N-1;\alpha)\sqrt{\frac{\tilde{V}}{N}} \tag{39}$$

と定められる。95%信頼区間（α=0.05）については、t(n;0.05)の値を表 1 に挙げている。

４．直線回帰と信頼区間

４．１　最小二乗法による回帰直線を求める

この節の説明は、１章冒頭の繰り返しから入る。図７はその説明図である。

ある量 y が x の関数として次式の線形関係で与えられている。

$$y = a + bx \tag{40}$$

x と y に関する母集団での関係と言ってよい。x が原因（説明変数）、y が結果（観測値、目的変数）と捕らえられるような量である。

実際に観測される y 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 ごとの値は、次式となる。

$$y_i = a + bx_i + e_i \tag{41}$$

以下では、第１章で示した４つ前提：i)～v）のもとでモデル化が行われる。重要なのでこれを再掲する。

i) 変数 x は誤差を含まない。（誤差は y 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 $N(0,\sigma^2)$の正規分布に従う（正規性）

x と y の組で与えられる N 個のデータ［$(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$］の処理によって得られる回帰直線

$$\hat{y} = \hat{a} + \hat{b}x \tag{42}$$

は、母集団の特性、すなわち、(40)式を推定するものになる。

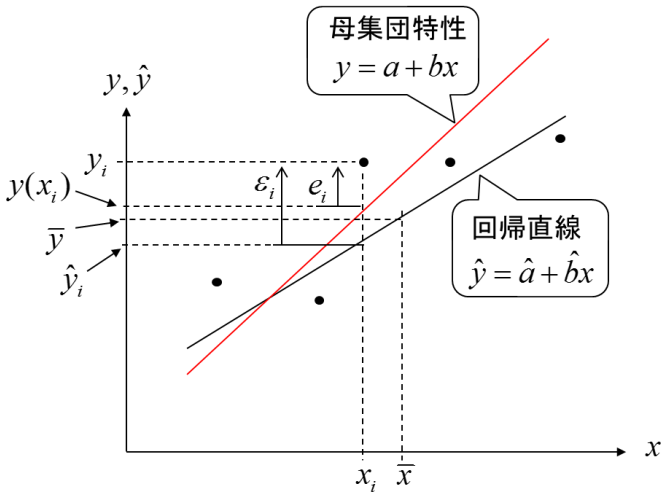


図７　データの母集団特性と回帰直線（回帰直線が重心 $(\bar{x}, \bar{y})$ を通ることは後に説明する）

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 2.1 節で述べた最小二乗法を用いる。最小二乗法では図 7 に示す残差 $\varepsilon_i = y_i - \hat{y}_i$ の 2 乗和

$$S_{\varepsilon\varepsilon} = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i)^2 \tag{43}$$

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、(41) 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。

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,,\ \ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{44a, b}$$

$$\bar{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \,,\quad \bar{y^2} = \frac{1}{N}\sum_{i=1}^{N} y_i^2 \tag{45a, b}$$

$$S_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2 = N(\bar{x^2} - \bar{x}^2), \quad S_{yy} = \sum_{i=1}^{N} (y_i - \bar{y})^2 = N(\bar{y^2} - \bar{y}^2) \tag{46a, b}$$

$$S_{xy} = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = N(\bar{xy} - \bar{x}\,\bar{y}) \tag{46c}$$

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。

$$\frac{\partial S_{\varepsilon\varepsilon}}{\partial \hat{a}} = \sum_{i=1}^{N} 2\varepsilon_i \frac{\partial \varepsilon_i}{\partial \hat{a}} = -2\sum_{i=1}^{N} \varepsilon_i = -2\sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i) = 0 \tag{47a}$$

$$\frac{\partial S_{\varepsilon\varepsilon}}{\partial \hat{b}} = \sum_{i=1}^{N} 2\varepsilon_i \frac{\partial \varepsilon_i}{\partial \hat{b}} = -2\sum_{i=1}^{N} \varepsilon_i x_i = -2\sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i) x_i = 0 \tag{47b}$$

上式は、以下の連立方程式に整理される。

$$\hat{a} + \hat{b}\bar{x} = \bar{y} \tag{48a}$$

$$N\hat{a}\bar{x} + (S_{xx} + N\bar{x^2})\hat{b} = S_{xy} + N\bar{x}\,\bar{y} \tag{48b}$$

これより、$\hat{a}, \hat{b}$ は以下のように定まる。

$$\hat{a} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \tag{49a}$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \tag{49b}$$

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、(47)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、(48a)式より重心$(\overline{x},\overline{y})$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 r とは

$$\hat{b}=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\sqrt{\frac{S_{yy}}{S_{xx}}}=\sqrt{\frac{S_{yy}}{S_{xx}}}\,r \tag{50}$$

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。

残差ε_iの二乗和 $S_{\varepsilon\varepsilon}$は

$$S_{\varepsilon\varepsilon}=\sum_{i=1}^{N}(y_i-\hat{a}-\hat{b}x_i)^2=S_{yy}-\frac{S_{xy}^2}{S_{xx}}=S_{yy}(1-r^2) \tag{51}$$

となり、相関係数と関係付けられる。相関係数の二乗値 r^2 は寄与率（あるいは決定係数）と呼ばれる。寄与率が高い（＝1 に近い）ほど、線形回帰がうまく言っていると解釈できる。

４．２　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。３章での議論と同様に、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

$$\hat{b}=\frac{S_{xy}}{S_{xx}}=\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})\frac{1}{S_{xx}}=\sum_{i=1}^{N}\frac{x_i-\overline{x}}{S_{xx}}(y_i-\overline{y}) \tag{52}$$

であるので、$y_i-\overline{y}$ 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。

$$V_{\hat{b}}=\sum_{i=1}^{N}\left(\frac{x_i-\overline{x}}{S_{xx}}\right)^2\sigma^2=\frac{\sigma^2}{S_{xx}} \tag{53}$$

$\hat{a}$ の分散$V_{\hat{a}}$も同様に

$$V_{\hat{a}}=V_{\overline{y}}+\overline{x}^2V_{\hat{b}}=\left(\frac{1}{N}+\frac{\overline{x}^2}{S_{xx}}\right)\sigma^2 \tag{54}$$

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、3章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、χ^2分布における自由度の理解が重要になる。

三つの量：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。

$$S_{yy} = \sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N} y_i^2 - N\bar{y}^2 \tag{55a}$$

$$S_{\hat{y}\hat{y}} = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{N} \hat{y}_i^2 - N\bar{y}^2 \tag{55b}$$

$$S_{\varepsilon\varepsilon} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i^2 - 2\sum_{i=1}^{N} y_i\hat{y}_i \tag{55c}$$

残差自乗和 $S_{\varepsilon\varepsilon}$の式(55c)右辺の第3項は

$$\sum_{i=1}^{N} y_i\hat{y}_i = \sum_{i=1}^{N}(\hat{y}_i + \varepsilon_i)\hat{y}_i = \sum_{i=1}^{N} \hat{y}_i^2 + \sum_{i=1}^{N} \varepsilon_i(\hat{a} + \hat{b}x_i) = \sum_{i=1}^{N} \hat{y}_i^2 \tag{56}$$

となる。最終辺は、前節で述べた最小二乗法の帰結である(47)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、

$$S_{\varepsilon\varepsilon} = \sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} \hat{y}_i^2 \tag{57}$$

となり、三つの量は、

$$S_{yy} = S_{\hat{y}\hat{y}} + S_{\varepsilon\varepsilon} \tag{58}$$

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χ^2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(53), (54)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。

$$\hat{V}_\varepsilon = S_{\varepsilon\varepsilon}/(N-2) \tag{59}$$

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

$$\hat{y} = \hat{a} + \hat{b}x = \bar{y} + \hat{b}(x - \bar{x}) \tag{60}$$

として、$\hat{y}$の分散$V_{\hat{b}}$は

$$V_{\hat{y}} = V_{\bar{y}} + (x-\bar{x})^2 V_{\hat{b}} = \left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} \right) \sigma^2 \tag{61}$$

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、

$$\hat{a} + \hat{b}x \pm t(N-2,\alpha) \sqrt{\left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} \right) \hat{V}_\varepsilon} \tag{61}$$

となる。この式が、本レポート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、(61)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。この目的に対しては、以下の方法によって、その幅が決定できる。

観測値y_iは$\hat{y}_i + \varepsilon_i$であるので、その分散は、

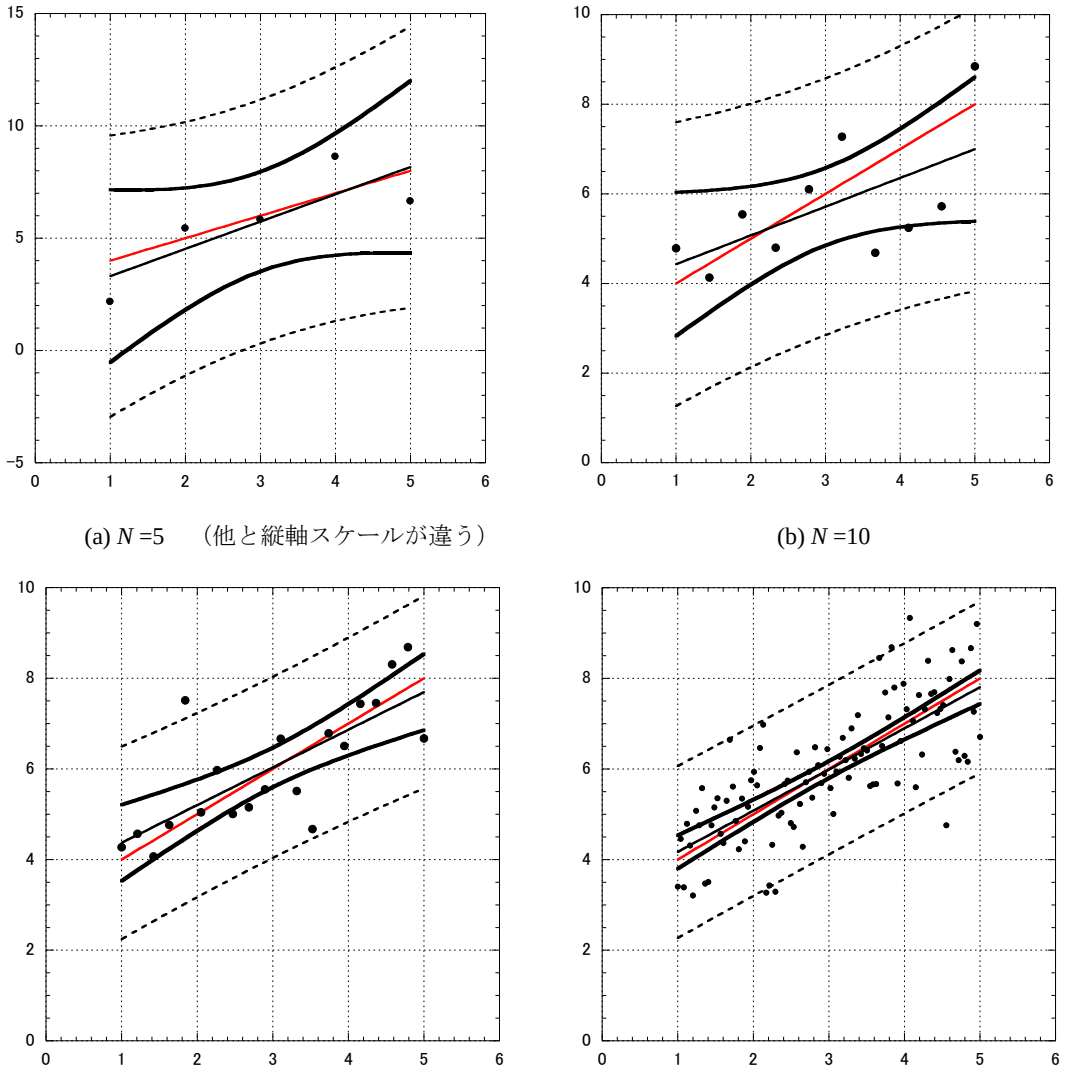
$$V_y = V_{\hat{y}} + V_\varepsilon = \left(1 + \frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} \right) \sigma^2 \tag{62}$$

これより、yの区間推定と同様に、y_iのばらつき幅（分布の信頼区間）の区間推定は、

$$\hat{a} + \hat{b}x \pm t(N-2,\alpha) \sqrt{\left(1 + \frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}} \right) \hat{V}_\varepsilon} \tag{63}$$

で求められる。この区間は個々の観測値のばらつきの範囲を表しているので、この範囲はNが増えても変化は少ない。Nの増加と共に、95%区間では$y = a + bx \pm 1.96\sigma$に近づいてくる。

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。図８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。



(a) N =5　（他と縦軸スケールが違う）　(b) N =10

(c) N =20　(d) N =100

図８　回帰直線と 95%信頼区間（グラフ：横軸 x, 縦軸 y、データ： a =3, b =1, σ =1; グラフ：赤実線：母集団特性、黒実線：回帰直線、太実線曲線：母集団特性の 95%信頼区間、外側の点線：データのばらつきの 95%信頼区間）

５．さらに理解を深めるために

５．１　モデル選択の重要性

本レポートでは、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（Wikipedia や文献[2]などに）。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。

図９は、アンスコムが提示した４つのケースの散布図と、それぞれの回帰直線と相関係数(*r*; 図では *R* で表記)を示している（元データの数値は[2]の表 4.5.1 より）。(a)は特段に問題なく、線形近似が妥当な例。(b)は曲線関係の有るデータに直線近似を行っていて、採用モデルが間違っている。(c)は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ている。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。(d)は回帰直線が、右端の一点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がある。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。

５．２　*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、１章の図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、本稿で述べた解析手法（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、*y* 方向ではなく、回帰直線と直交する方向に見る）よう変更しなければならず、その場合の回帰直線は図に示した 45°の線になる。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。

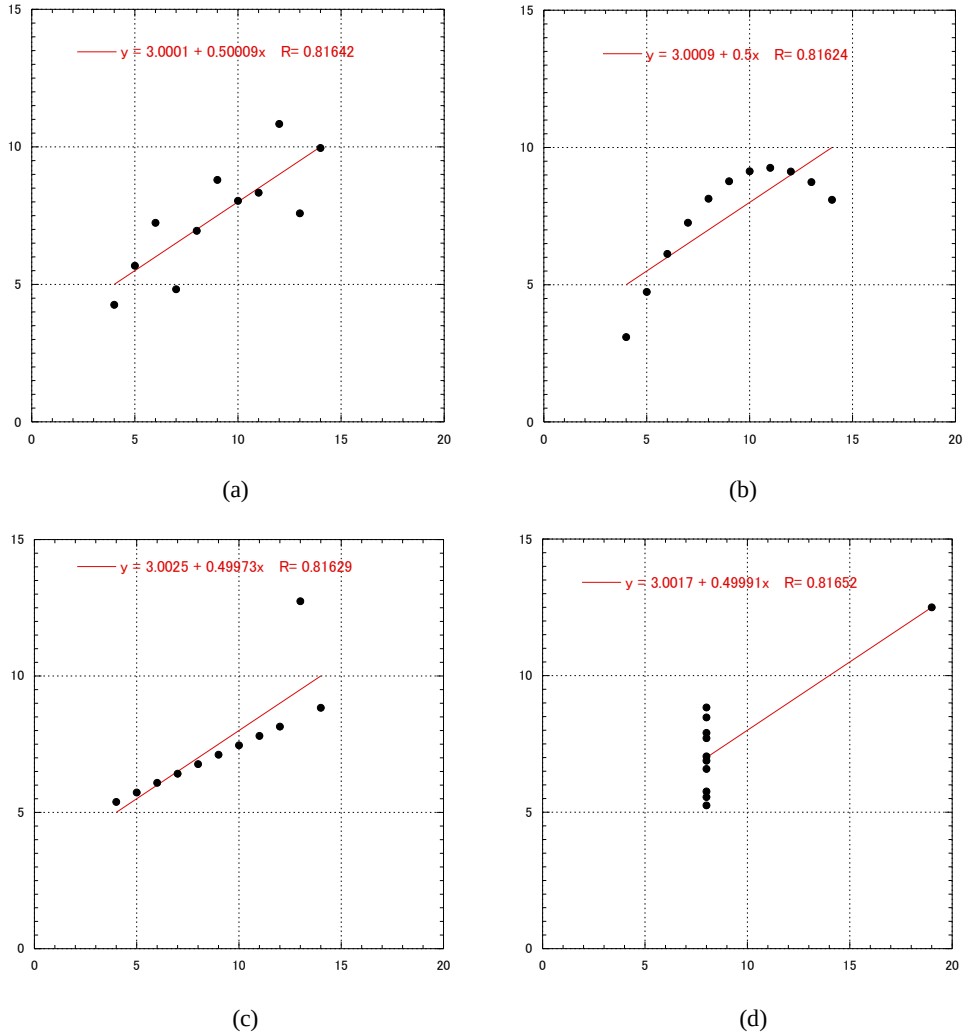


図９　アンスコムの例（平均・標準偏差・相関係数が等しく、かつ、回帰直線も同じになる４つの例。散布図で確認すると違いが歴然）

５．３　参考文献等

本資料では、回帰分析のもっと基本である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。

筆者はこの資料作成のために、以下の本を参考にしている。説明や数式の展開に大いに影響を受けている。

[1] 前園宣彦, 概説 確率統計[第３版], サイエンス社, 2018.10.
（入門編として回帰分析の基本を学ぶのにお勧め。*t* 分布（や *F* 分布）の数値表が載っていて便利）

[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.（応用分野は違うが統計解析の基礎は同じ。直感的理解を呼び起こすユニークな図や記述が多い）

[3] 芝 祐順他, 統計用語辞典, 新曜社, 1984.
（確率統計の勉強をする上において、用語の知識を得るに大いに役立つ）

次のステップとして、多変量解析に進むには、入門から本格的なものまで多くの専門書が出版されているので、身の丈にあったものを選んでほしい。後者に位置づけられるのもとして[4]を挙げておく。[4]では、本資料内容の単回帰モデルがベースとなる多変量の重回帰分析や、異なる視点でデータ分析を行う主成分分析について、物理的意味を追いながら丁寧にまとめられている。

[4] 奥野忠一他, 多変量解析法 改訂版, 日科技連出版社, 1981.

統計の理解にはそのベースになる確率論やその帰結である種々の確率分布の勉強も必要である。確率論に関しても、多くの教科書・専門書が出版されているので、本屋さん等で手にとって見て、相性のよさそうなものを自分で選んでほしい。筆者も無線通信に現れる確率分布を以下にまとめているので、必要なときに参考にしてほしい。

[5] 唐沢好男, 改訂 ディジタル移動通信の電波伝搬基礎（第３章）, コロナ社, 2016.

[6] 唐沢好男, “伝搬モデルに現れる確率分布：～レイリーフェージングから Massive MIMO まで～,” 私報、Tech. Rep. YK-005, pp. 1-36, Dec. 2017.
http://www.radio3.ee.uec.ac.jp/ronbun/TR_YK_005_Probability_distribution.pdf