

zスコア理論編

一般社団法人 日本環境測定分析協会
技能試験委員会

1. 解説のための基礎

確率変数 X の分布関数を $F(x)$ とする。即ち、確率変数 X が x を超えない確率を $F(x)$ とする。この分布関数 $F(x)$ が少なくとも2つのパラメータ μ と σ を含むとしたとき、

$X - \mu$ の分布関数は μ を含まない

ならば μ を X の位置パラメータといい、更に

$\frac{X - \mu}{\sigma}$ の分布関数は σ を含まない

ならば σ を X の尺度パラメータという。

ここでの記号 X , x , μ 及び σ などは量を表す。量とは数と単位の組のことをいう。分布関数 $F(x)$ は確率であるから数である。

確率変数 X に対して位置パラメータ μ と尺度パラメータ σ を用いて

$$U = \frac{X - \mu}{\sigma}$$

の如く変換することを「標準化変換」という。これは量 X から数 U への一つの変換であり、 X の位置パラメータと尺度パラメータとに依存しなくなる変換であることを意味するのみならず、同時に如何なる単位の量 X についてもこの標準化変換がなされた後の数 U に於いて、大小の比較などが互いに可能であることをも意味する。

1つの確率変数 X の n 個の実現値を $(x_1, x_2, x_3, \dots, x_n)$ とする。これは同一の母集団から無作為に大きさ n の標本を採取することに対応する。一方でこの大きさ n の標本 $(x_1, x_2, x_3, \dots, x_n)$ は、互いに独立に同一の分布に従う n 個の確

率変量の組 $(X_1, X_2, X_3, \dots, X_n)$ の一つの実現値であるとも見なされる。この $(X_1, X_2, X_3, \dots, X_n)$ は標本確率変量とも呼ばれ、統計量である。標本 $(x_1, x_2, x_3, \dots, x_n)$ をどちらの意味での実現値として解釈するかは任意である。

大きさ n の標本 $(x_1, x_2, x_3, \dots, x_n)$ を小さい順に並べ替えた後に、新たに番号を付け替えた

$$x(1) \leq x(2) \leq x(3) \leq \dots \leq x(n)$$

を満たす組

$$(x(1), x(2), x(3), \dots, x(n))$$

が実現値である確率変量の組

$$(X_{(1)}, X_{(2)}, X_{(3)}, \dots, X_{(n)})$$

を順序統計量という。例えば $x(1)$ は標本最小値であるけれども、この性質は $X_{(1)}$ の分布を調べることによって得られる。

分布関数に対応するものとして、経験分布関数

$$\hat{F}(x) = 0, \quad x < x(1)$$

$$\hat{F}(x) = \frac{i}{n+1}, \quad x(i) \leq x < x(i+1), \quad i = 1, 2, 3, \dots, n-1$$

$$\hat{F}(x) = 1, \quad x(n) \leq x$$

を定義する。これは階段状の非減少関数である。(図1)

分布関数 $F(x)$ についての方程式

$$F(x) = \frac{i}{4}, \quad i = 1, 2, 3, 4$$

を満たす解 x がただ一つ存在するとき(一般には一つに限らないけれどもここでは一つに限ることにする)、これを確率変数 X の第 i 四分位値という。(図2) 特に $i=2$ の場合は方程式 $F(x) = 0.5$ のときであるから、確率の意味でちょうど真ん中を与える値であることから、確率変数 X の第2四分位値は母中央値とも呼ばれるものであり、確率変数 X の一つの平均値を定義している。また、第3四分位値と第1四分位値との差は分布の拡がりの大きさを示す指標の一つとなる。

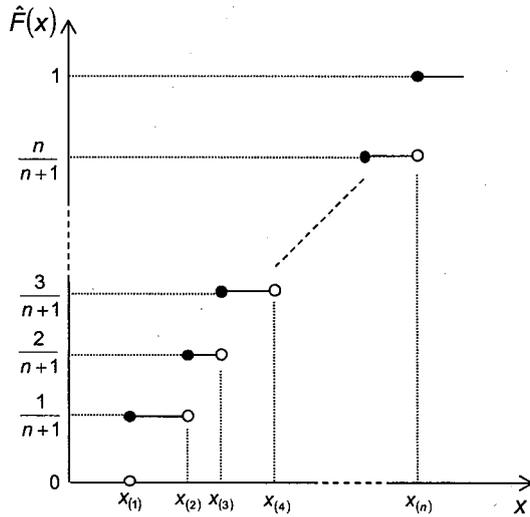


図 1 経験分布関数

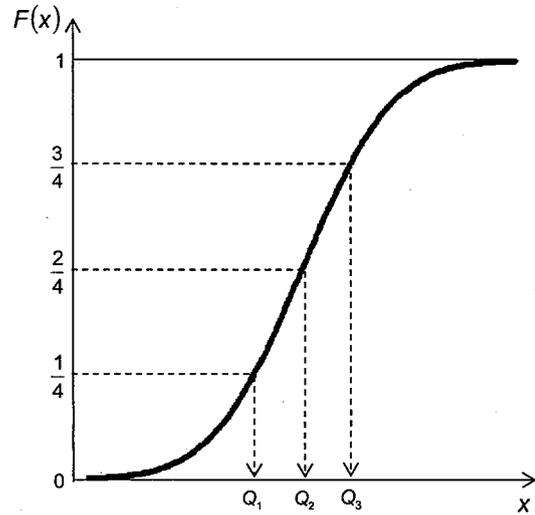


図 2 分布関数と四分位数

分布関数に対応する経験分布関数でも同様な考え方で標本第 1 四分位値 Q_1 、標本第 2 四分位値 Q_2 及び標本第 3 四分位値 Q_3 が定義されるけれども、経験分布関数は不連続な階段関数なので、該当する階段部分を単純に線形補間することにより求めることが多い。

2. 正規分布を前提とした z スコアによる評価

確率変数 X が正規分布に従うとは、その分布関数 $F(x)$ が

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] \frac{dt}{\sigma}, \quad -\infty < x < \infty$$

であるときをいい、記号で $X \sim N(\mu, \sigma^2)$ と記される。量であるパラメータ μ および σ の存在範囲はそれぞれ

$$-\infty < \mu < \infty, \quad 0 < \sigma < \infty$$

である。なお、ここでは指数関数 e^x のことを $\exp(x)$ と書いたものであり、 e とは自然対数の底である。

確率変数 X が正規分布 $N(\mu, \sigma^2)$ に従うとき、 X の位置パラメータ μ は X の母中央値に等しいのみならず母算術平均値（以下、混乱がないときは平均値と略）にも等しい。また、 X の尺度パラメータ σ は X の母標準偏差に等しい。これらパラメータで標準化変換された確率変数 $U = (X - \mu)/\sigma$ は標準正規分布

$$U \sim N(0, 1^2)$$

に従う。

平均値である μ の推定量としては沢山あるけれども、ここでは標本第 2 四分位値 Q_2 を採用する。また、標準偏差である σ の推定量も沢山あるけれども、ここでは四分位値を使った $0.7413 \times (\text{第 3 四分位値 } Q_3 - \text{第 1 四分位値 } Q_1)$ を採用する。これらを代入して得られる数

$$z_i = \frac{x_i - Q_2}{0.7413(Q_3 - Q_1)}, \quad i = 1, 2, 3, \dots, n$$

を「 z スコア」という。標準化変換の所でも述べたように、この z スコアは数であり、元のデータ x_i の単位が如何なるものであろうとそれら単位に依存しない数に帰着されているものである。従って、相異なる単位のデータから計算された z スコアは互いに比較可能である。しかし残念ながら、 z_i が標準正規分布に従う確率変数の実現値であるという性質は失われている。

z スコア z_i は標準正規分布に従う確率変数の実現値ではないけれども、概ねそれに近い分布に従う確率変数の実現値であると見なして、以下のように判断する。

$$\begin{aligned} 0 \leq |z_i| \leq 2 &\Rightarrow \text{満足} \\ 2 < |z_i| < 3 &\Rightarrow \text{疑わしい} \\ 3 \leq |z_i| < \infty &\Rightarrow \text{不満足} \end{aligned}$$

【参考】確率変数 U が標準正規分布 $N(0, 1^2)$ に従うならば、 $0 \leq |U| \leq 2$ を満たす確率は 0.9544 であり、 $2 < |U| < 3$ を満たす確率は 0.0430 であり、 $3 \leq |U| < \infty$ を満たす確率は 0.0026 である。

上記の z_i による判断では以下のことに注意をすべきである。

1. データの相対的な位置関係を述べているに過ぎないものである。毎年蓄積される自社の z スコアデータを様々な統計的方法、例えば QC 七つ道具などを用いて解析し改善活動などに活用すべきである。
2. 必ずしも正規分布を想定できるとは限らない場合も多くあり、従って正規分布の仮定に基づく判定基準は参考の目処に過ぎない。
3. 正規分布を想定できたとしても、パラメータの推定量としては他に多くの推定量がある中で、情報の損失はあるにせよ計算の簡便さなどを重視して四分位値を用いているものである。言い換えれば、四分位値を用いない z スコアも当然のことながら定義できる。

3. 2 試料に基づく z スコアによる評価

各 i 試験所は互いに濃度が多少異なる二つの試験試料 A, B を測定する。即ち、二つの試料ともに同じ単位で測定されることを前提とする。その測定結果を記号 (x_{Ai}, x_{Bi}) とすると、 n 箇所の試験所全部から集められたデータは

$$(x_{A1}, x_{B1}), (x_{A2}, x_{B2}), (x_{A3}, x_{B3}), \dots, (x_{An}, x_{Bn})$$

となる。

これらのデータを、縦軸横軸ともに同じスケールの直角座標に打点すると、第一象限部分に打点され、更にそれらの点は原点を通る直線に沿ってばらついた状態になっている。

この図での想定される原点を通る直線の傾きは二つの試験試料の濃度の相違で決まる。試験試料 A の方が B よりも濃度が高ければ、原点を通る直線の傾きは角度が $\pi/4$ よりも小さくなるし、その逆の場合には角度は $\pi/4$ よりも大きくなる。現実には二つの試験試料の濃度はあまり相違がないので、角度はほぼ $\pi/4$ 前後である。

以下で試験試料 B の中央値（第 2 四分位値 Q_2 ）の方が試験試料 A の中央値よりも小さい場合を仮定して話を進める。

4. 想定される点の性質

仮に各試験所内で測定のばらつきが無いとすれば、二つの試験試料の濃度の比は正確に反映され、各試験所間のかたよりが反映されて打点されることになる。このような状況の下では、図3に示すように、原点を通る1本の直線の上にそれぞれの試験所の結果が散らばって打点されていることになる。

一方、各試験所間のかたよりに及び各試験所に配付される試験試料のばらつきがないとすれば、各試験所内でのばらつきだけが反映されて打点されることになる。このような状況の下では、試料 A、試料 B の両方に同じだけ測定のばらつきがあるから、図4に示す通り、試料 A の真の濃度と、試料 B の真の濃度の点を中心とする円状に打点されて、図3の原点を通る直線と直交する方向にもばらついて点が存在するようになる。

以上のように、両極端であるような二つの場合を想定して考えることにより、想定される直線方向での点のばらつきと、その直線に直交する方向での点のばらつきとを分けて考察することにより、各試験所間のかたよりと各試験所内でのばらつきの情報を汲み取ることができるとなる。(図5)

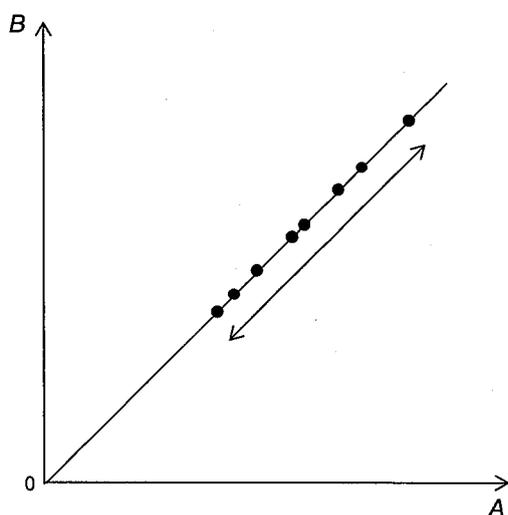


図3 かたよりの概念図

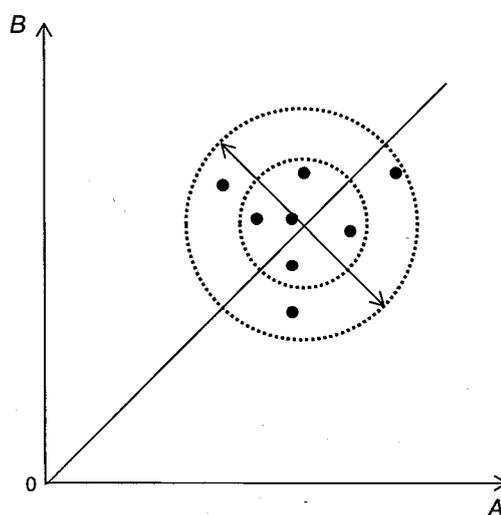


図4 ばらつきの概念図

尚、これ以降の考察の前提条件は、二つの試料の濃度に過度な相違がないことと、配付される試験試料のばらつきは各試験所間のかたよりの大きさとか試験所内のばらつきとかと比較して十分に小さく無視し得ることである。更に非標本誤差は考慮の対象外である。実際には計測時の読み取りの桁違いやデータの転記ミスなどと考えられる場合が少なからず存在するからであり、これら

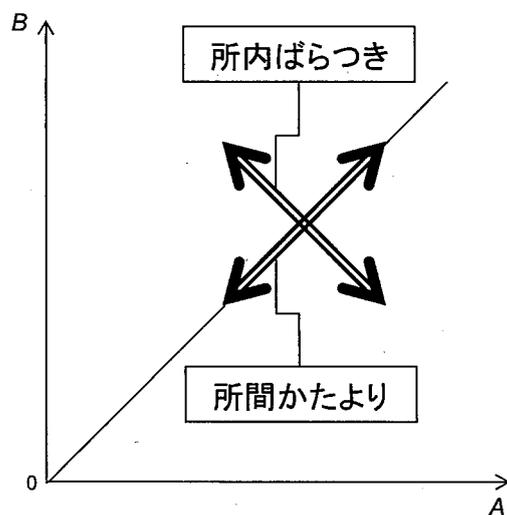


図5 所間かたよりと所内ばらつき

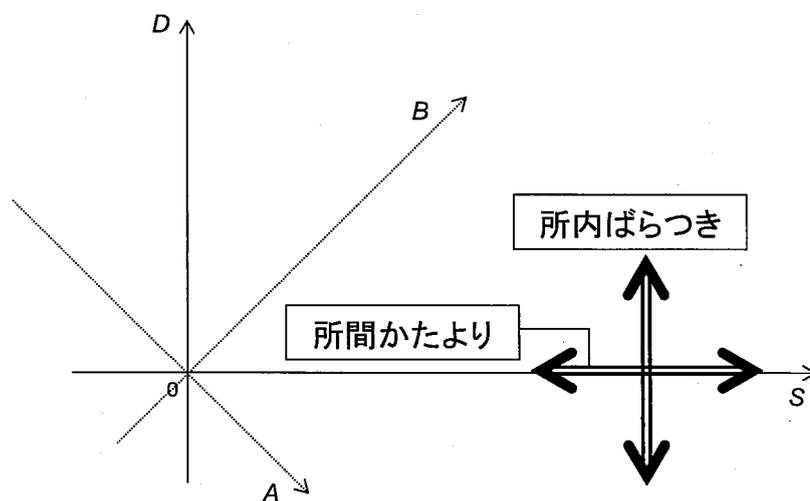


図6 回転後の所間かたよりと所内ばらつき

は固有技術に基づく考察によって事前に排除されていなければならない。各試験所間のかたよりの情報と各試験所内でのばらつきの情報を見るために、横軸 A 、縦軸 B の座標を $\pi/4$ だけ負の方向へ（時計回り方向へ）回転して考察することによって、新しい横座標は概ね各試験所間のかたよりの情報を与え、新しい縦座標は試験所内でのばらつきの情報をあたえることとなる。（図 6）

5. 座標の回転に基づく z スコア

$A-B$ 座標軸系での点 (A_i, B_i) は、この $A-B$ 座標軸を $\pi/4$ だけ負の方向へ回転させた新しい座標軸系 $S-D$ で表現すると

$$S_i = A_i \cos \frac{\pi}{4} + B_i \sin \frac{\pi}{4} = \frac{A_i + B_i}{\sqrt{2}}$$

$$D_i = -A_i \sin \frac{\pi}{4} + B_i \cos \frac{\pi}{4} = \frac{-A_i + B_i}{\sqrt{2}}$$

となる。

例えば、 $A-B$ 座標系での点 $(1, 1)$ は、 $A_i = 1, B_i = 1$ を上式に代入すれば、 $S_i = \sqrt{2}, D_i = 0$ となるから、 $S-D$ 座標系では $(\sqrt{2}, 0)$ となる。（図 7）

[注意] 元の座標系である $A-B$ 座標系を角度 θ だけ回転させた新しい座標系を $S-D$ 座標系とする。 $A-B$ 座標系での点 (A_i, B_i) を新しい $S-D$ 座標系から見た座標 (S_i, D_i) は

$$S_i = A_i \cos \theta - B_i \sin \theta$$

$$D_i = A_i \sin \theta + B_i \cos \theta$$

によって表されるという関係を使ったに過ぎない。 z スコア法では

$$\theta = -\frac{\pi}{4}$$

と定めている。これは二つの試験試料の濃度にあまり相違がないことを前提としてのことである。しかしながら、実際は濃度にある程度の相違があり

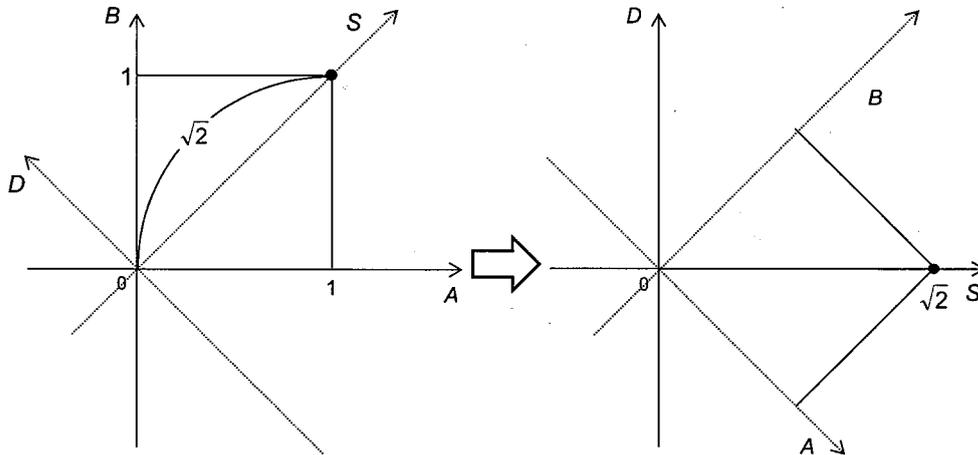


図7 座標の回転例

$\theta = -\pi/4$ とすることが望ましくない場合も散見される。

この新しい座標系での各試験所からのデータによる点の様子を見ると、先に述べたように、 A_i と B_i とは近い濃度だから、従って D_i は零の近辺をばらつくことになり、この値の正負はそのばらつき方向を示していることになる。この値が負の値になることは全体の約半分程度はあることになるが、所内ばらつきの大きさについて、全体の中で相対的に正の方向に大きい小さいか、負の方向に大きい小さいかを表していることになる。しかしながら、元々の A_i と B_i との濃度差も反映されているものであり、正負が約半分ではなくて、正または負の割合が多くなることもある。そして、この全体でのばらつきは「所内ばらつきのばらつき」を表していることになる。この値の絶対値が小さければ小さいほど、相対的にその試験所内でのばらつきが小さいということになる。また、 S_i は各試験所間のかたよりの大きさを表していることになる。

実際に打点された例を見ると、想定された直線よりも上側に点が多く存在しているような場合が多いけれども、これは単に二つの試験試料で B_i の方が A_i よりも濃度が高いだけのことであり、 D_i が全体として正である割合が多いとか負である割合が多いとかは本質的なことではない。

S_i 、 D_i は元の測定単位と同じ単位を有した量である。また、 D_i の解釈のとこ

るでも述べたように、元の A_i と B_i との濃度差も反映されている。これらから解放されるために標準化変換の一つである z スコア化を S_i, D_i に対して施す。

このデータより z スコアを計算したものは

$$(z_{S_1}, z_{D_1}), (z_{S_2}, z_{D_2}), (z_{S_3}, z_{D_3}), \dots, (z_{S_n}, z_{D_n})$$

として得られることになる。

上記 z スコアを直交座標に打点し、判定基準の ± 2 および ± 3 で囲まれた正方形枠と、各試験所のこの z スコア点の位置を見ることによって、各試験所のばらつきとかたよりの全体の中での相対的な能力を確認することが可能となる。

尚、この標準化された z スコアの座標系では判定枠は正方形であるが、これらを全て逆変換して、元の $A-B$ 座標系に判定枠を描いたものが多く用いられ、後述の信頼域楕円と比較される場合がある。

6. 信頼域楕円との関係

3章で示したような2種類の試料について得られた (x_{Ai}, x_{Bi}) を打点した散布図はユーデンプロットと呼ばれる。ユーデンプロットの解析には z スコアに基づく判定枠だけではなく、信頼域楕円が用いられることもある。信頼域楕円による解析は2変量正規分布に基づいている。2変量正規分布 $N(\mu_A, \mu_B; \sigma^2_A, \sigma^2_B; \rho)$ の確率素分は

$$f(x, y) dx dy = \frac{1}{2\pi} \frac{1}{\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\left(\frac{x-\mu_A}{\sigma_A} \right)^2 - 2\rho \frac{x-\mu_A}{\sigma_A} \cdot \frac{y-\mu_B}{\sigma_B} + \left(\frac{y-\mu_B}{\sigma_B} \right)^2 \right) \right] \frac{dx dy}{\sigma_A \sigma_B}$$

$$-\infty < x < \infty, -\infty < y < \infty$$

で与えられる。ここで、 (x, y) の曲線の方程式

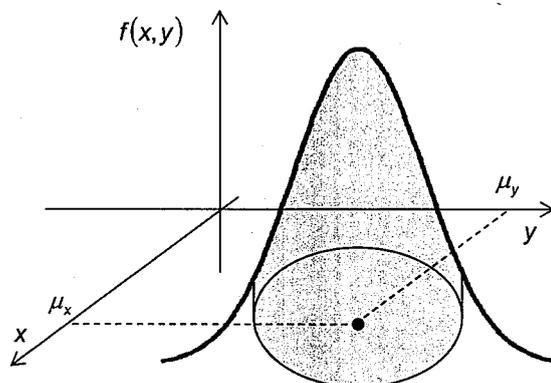


図8 2変量正規分布

$$\frac{1}{1-\rho^2} \left(\left(\frac{x-\mu_A}{\sigma_A} \right)^2 - 2\rho \frac{x-\mu_A}{\sigma_A} \cdot \frac{y-\mu_B}{\sigma_B} + \left(\frac{y-\mu_B}{\sigma_B} \right)^2 \right) = k^2, \quad 0 < k < \infty$$

を考え、この曲線で決められる楕円柱部分の体積が 0.95 または 0.99 であるような k に基づき、それぞれのパラメータの推定値が代入されて描かれたのが信頼域楕円と考えて差し支えない。(図 8)

ここで、母相関係数 ρ の意味するところは、各点が想定される直線にどの程度集中しているかどうかの指標の役割を果たしている。 $|\rho|$ が 1 に近いと、長軸半径と短軸半径が極めて相違のある細長い楕円となる。

この楕円の長軸と短軸を軸とするような座標軸に変換をすれば、当初 z スコアが狙ったような情報を検出できることになる。

【参考】正規分布 $N(\mu, \sigma^2)$ の確率素分

$$f(x)dx = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \frac{dx}{\sigma}, \quad -\infty < x < \infty$$

において、(座標軸の $\pi/4$ 回転後に) 試料 A, B についてそれぞれ別々に

$$\left(\frac{x-\mu}{\sigma}\right)^2 = k^2, \quad 0 < k < \infty$$

を満たす二つの x の値で挟まれる面積が 0.9544 または 0.9974 であるような k に基づき（実はそのような k はそれぞれ 2 および 3 である）描くのが 5 章に説明した z スコア法に基づく判定枠であると考えて差し支えない。（図 9）

信頼域楕円は (A_i, B_i) を 2 変量正規分布に従う確率変量の標本と見なして、分散共分散行列に基づいて描かれた楕円である。一方、 z スコアに基づく矩形領域では（暗に試料 A と試料 B の濃度が同じであると仮定して） $\pi/4$ の軸の回転を行った後に、データを両軸に射影し、その射影された両軸の値に対してそれぞれ単独に z スコアを算出し、元の座標に戻したものである。従って、理論的には信頼域楕円と、ここで述べた z スコアによる判定枠とを直接に結びつける関係はない。さらに、信頼域楕円が第一象限に収まらないような不自然な事象も理論的には起こりうる。従って、信頼域楕円による判定の方が z スコア法での判定よりも厳しくなる傾向はあるにしても一般論としては言えないことである。

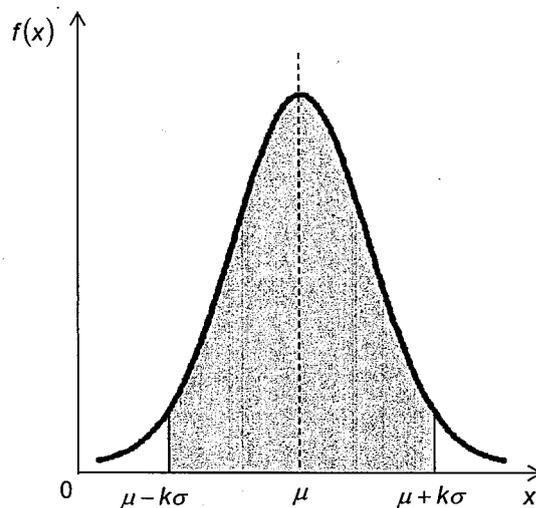


図 9 正規分布

しかしながら、現実には結果としてかなり近接した関係が得られることも事実である。両者の理論的な相違を認識した上で、両者による判定を比較すべきである。

信頼域楕円と z スコア法に基づく判定枠とを直接比較しようとするならば、座標軸の回転を $\pi/4$ などのように機械的に決めてしまうのではなくて、データよりその回転角を求めて回転し（即ち、信頼域楕円の長軸および短軸に一致するように座標の回転をし）、この下での (S_i, D_i) に対して z スコア法を適用し、判定枠を定めることにより達成される。（図 10）

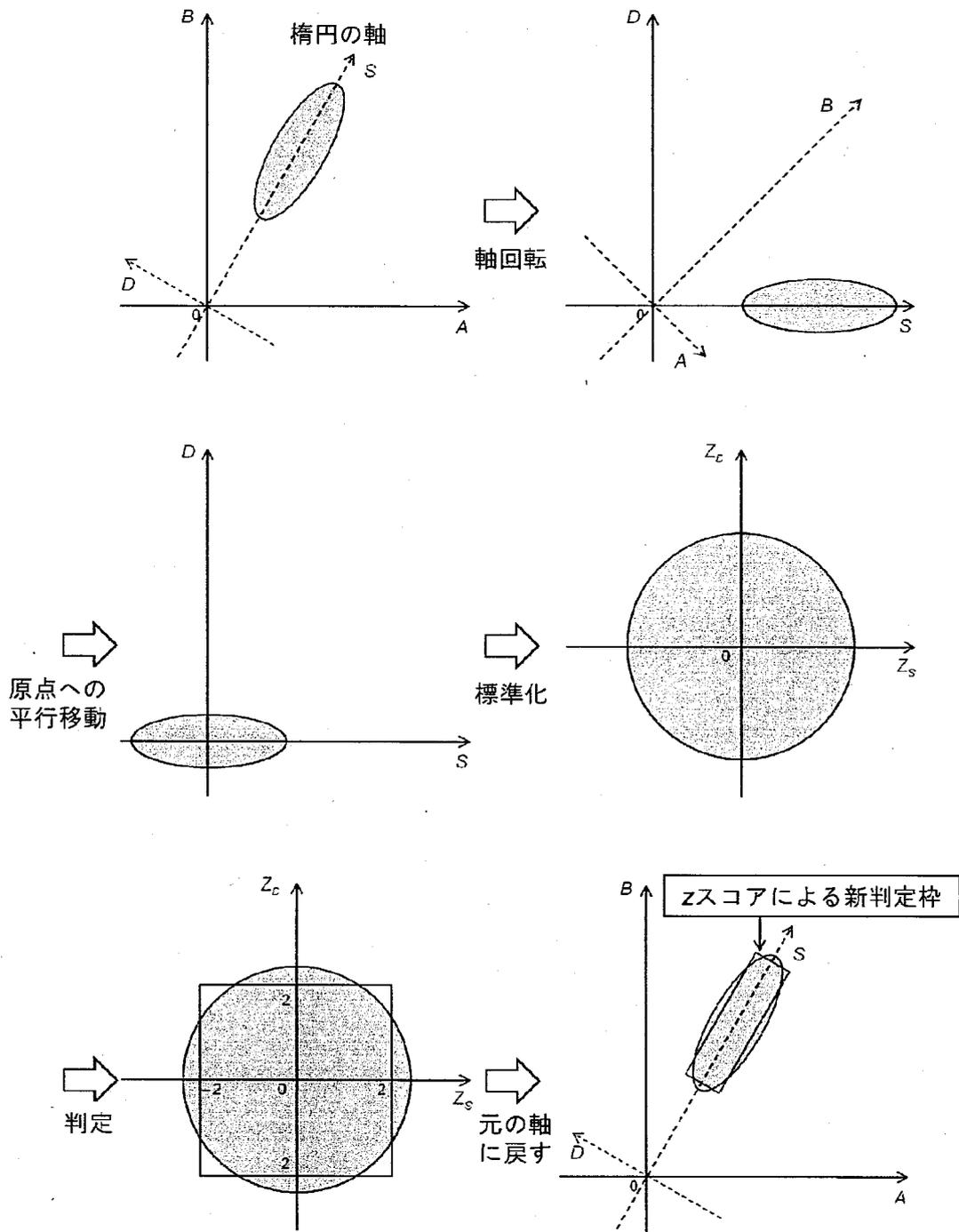


図 10 座標変換と新しい判定枠の関係