測定における統計解析の基礎

田　中　秀　幸

この度，2024 年の入門講座として「データ解析：定量・定性からビッグデータの解析まで」を企画いたしました．

近年の社会環境の変化に伴い，分析技術は高精度化・高感度が求められ，一方で分析対象物は高精密化・高機能化に伴い複雑化しております．このような分析対象物から有用な情報を得るためには，データ解析は重要であり，日々多くの分析化学者に活用されております．しかしながら，簡単に使用できる解析ソフトウェアにより，その原理まで深く理解せずともそれを使用できます．さらに得られたビッグデータを統計解析することまで可能となりました．

本入門講座では，「データ解析：定量・定性からビッグデータの解析まで」と題しまして 12 個のテーマを取り上げ，代表的な分析機器における定性・定量などのデータ解析，統計解析，さらにビッグデータの統計解析事例についてご執筆いただきました．分析化学者が普段何気なく活用しているデータ解析を改めて深く理解していただくことで，さらなる活用のきっかけとなれば幸いです．

〔「ぶんせき」編集委員会〕

1　はじめに

本解説では，測定データに適用する統計的手法について，もっとも基礎となる標本・母集団についての考え方から，母平均の区間推定の考え方の基礎までを解説する．

特に，標本・母集団の考え方は統計的手法の根本となる考え方であり，この標本・母集団に関する統計的視点を測定データに適用することができるようになると初めて測定データが意味のあるものとして見えてくるかと思う．統計について苦手意識を持っている方に特に参考にしていただきたい．

ただし紙面に限りがあるため，基本的な統計を理解するための必要最小限の内容のみピックアップして解説する．ある程度統計が詳しい読者の方からすると，重要な統計的知識を全く説明していない，と感じる解説かもしれない．ただ今回は分かりやすさをとにかく重要視し，大胆に教えるべき内容をカットしていることをまずもって断っておく．

2　標本と母集団

測定における統計を考えるうえで一番基本となる考え方が標本と母集団である．測定における標本とは，実際に測定したデータのことである．また母集団とは，その測定を無限回行ったとしたときに考えることができる分布のことである．標本を数多く集めていくと母集団に関する情報が集まっていき，無限回測定を行ったとすると完全に母集団を知ることができる，ということである．

次にこれを逆転して考えてみよう．つまり，測定を行おうと考えた時には（我々には知ることができないが）何らかの母集団が決定している．そして，測定を行えば，その母集団からデータをランダムにサンプリングして得ることができる，と考えられるだろう．その「母集団からのサンプリング」という行為が統計的視点から見た測定の本質である．

次に，測定を何回か繰返していくつかのデータを得た後に，そのデータの平均値を求めて，それをその測定の代表値とする，ということを考えてみよう．このとき測定者は何を知りたいのだろうか？取得したデータの平均値が知りたいのだろうか？それは違うだろう．本当に知りたいのはその測定の真の平均値だろう．つまり，無限回測定を繰返したときに得られる母集団の平均値（母平均）を本当は知りたいのだが，しかし無限回の測定を行うことはできない．よって有限回の測定の繰返しを行い，そこで得られたデータの平均値（標本平均）を求めることによって，それを母平均の代用として使用している，ということを行っている．つまり，標本平均によって母平均を推定している，ということである．

しかし，標本平均によって母平均を推定したとしてもその推定がどの程度うまくできているのか，ということはよく分からない．例えば，母集団のばらつきが大きい場合，そこから標本を取り出して標本平均を求めるという操作を何回も行ったとすると，その求められた標本平均は様々な値をとってしまうだろう．逆に母集団のばらつきが小さい場合には，求められた標本平均はだいたい同じくらいの値をとることが予想できるだろう．よって母平均の推定値だけをそのまま提示しただけであれば，

The Fundamentals of Statistical Analysis in Measurement.

その値がどのくらい信用できるのか，ということがよく分からない．つまり推定した値がどの程度信頼できるのか，ということを求めることができれば，それは非常に有用な情報提供となる．

それでは，推定値がどの程度信頼できるのか，ということを表すためには何が必要かを考えてみよう．前述したように，いくつかのデータから標本平均を求める，という操作を何回か繰返した場合，得られる何個かの標本平均間のばらつきは母集団のばらつきが反映されるだろう．つまり母平均の推定値である標本平均がどの程度信用できるか，ということは元の測定の母集団のばらつきの大きさに依存するということである．しかし，測定の母集団のばらつきの大きさは母平均と同じく標本から推定せざるを得ない．通常統計ではばらつきは「標準偏差」もしくは標準偏差の二乗に相当する「分散」によって表される．「標準偏差」とは簡単にいうと「ばらつきの平均値」である．そして標準偏差，分散にも母集団のもの，標本のものがあり，それぞれ，「母標準偏差」「母分散」「標本標準偏差」「標本分散」と呼ばれる．また，母集団の性質を表す「母平均」「母分散」「母標準偏差」等は「母数」と呼ばれ，また「標本平均」「標本分散」「標本標準偏差」等は「統計量」と呼ばれる．

まとめると，測定の目的は，

・測定を代表する値として用いる標本平均
・測定結果がどの程度信用できるかを求めるための母集団のばらつきの推定値

の上記二つを求めることである．

それでは「測定結果がどの程度信用できるか」ということを客観的に示す方法はどのようなものがあるだろうか？

例えば，よく行われている測定結果の提示例を思い出してみると，

測定結果
溶液中に含まれる Cd^{2+} イオンの濃度
1.23 mg/L ± 0.13 mg/L

のように，正しい値が含まれるであろう範囲として表すことが多い．よってここからは，測定を代表する値と測定データのばらつきを求め，そこから正しい値が含まれるであろう範囲を求める手法を解説する．

3　分散と標準偏差

ここではばらつきの大きさを定量化することを考える．ばらつきは標準偏差や分散によって表されるが，それぞれの算出について解説する．

標準偏差とはいわば，ばらつきの平均値である，と説明した．これについて考えよう．ばらつきの平均値，ということはまずそれぞれの測定データのばらつきの大きさを知ることが必要である．それぞれのデータのばらつきの大きさは，標本平均からそれぞれのデータがどの程度離れているか，ということに相当する．これを式で表すことを考える．まず，測定値を表す変数を x とする．この x は測定を行うたびに母集団からサンプリングされて得られる値であり，データを取得するたびに値が変わる．このような変数を「確率変数」と呼ぶ．このとき繰返し測定を行い，データを n 個取得したとしよう．この場合あるデータが持つばらつきの大きさ（残差[*1]）は以下の式で表すことができる．

$$x_i - \bar{x} \quad \cdots\cdots (1)$$

ここに，$x_i (i = 1, ..., n)$ は測定データ，$\bar{x}$ は標本平均を表す．

次に平均的なばらつきの大きさを得るためにこの残差の和を求めるが，このまま残差の和をとっても 0 となるのは自明だろう．これは標本平均を中心にプラス方向にもマイナス方向にもデータがばらついているためである．よって残差を正の値とするためにすべての残差を二乗しその和を求める．

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \cdots\cdots (2)$$

これを残差の二乗和と呼ぶ．次に残差の二乗平均を求めるために残差の二乗和をデータの個数から 1 を引いた $n-1$ で割る．

$$s^2(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \cdots\cdots (3)$$

ここで求められた $s^2(x)$ のことを「標本分散」と呼ぶ．それではなぜ，この標本分散を求める際にデータの個数である n で割るのではなく $n-1$ で割るのだろうか？　これを考える上で重要なのは，ここで求めようとしているのは残差の二乗平均であるが，本当に意味のある残差はいくつあるか，ということである．確かに残差は n 個求まっているが，残差は式（1）で求められているため，先ほども触れたように残差の和は 0 になる．そうなると例えば，5 番目のデータ x_5 が分からなくなってしまったとしても，残差の全和が 0 となるので，

$$(x_5 - \bar{x}) = -\sum_{\substack{i=1 \\ i\neq 5}}^{n}(x_i - \bar{x}) \quad \cdots\cdots (4)$$

が成立する．つまり，残差が一つ分からなくなっても，分かっている残差の和を求めてそれにマイナスを付ければ分からなくなった残差を求めることができる，ということである．これはつまり本当に意味のある残差は

[*1] 「残差」ではなく「偏差」という用語を使う場合も多いが，本項では「残差」を「測定値と標本平均との差」，「偏差」を「測定値と母平均との差」の意味で用いる．

$n-1$ 個であるということを示している．この自由度について本稿ではこの程度の解説にとどめておくが，詳細を知りたい場合は「分散の不偏推定量」について調べてほしい．

さてここで求められた標本分散は残差の二乗平均を表しており，これは標本のばらつきを表すパラメータとして十分用いることができるものである．しかし標本分散は残差を二乗していることから，単位が元の測定データの単位の二乗となっており，測定データとは異なる次元で表されている．そうすると例えば，標本平均の大きさとばらつきの大きさを比較したい，というような場合，分散は異なる次元を持っているので比較ができないという問題が起こる．よって，ばらつきを表すパラメータで測定データと同じ次元を持つものを求める必要がある．ここで，標本分散の次元は測定データの次元の二乗となっているのであれば，標本分散の平方根を求めれば同じ次元のばらつきを表すパラメータを求めることができるだろう．

$$s(x)=\sqrt{s^2(x)}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \quad \cdots\cdots (5)$$

ここで求められた $s(x)$ のことを「標本標準偏差」と呼ぶ．

「標本分散」「標本標準偏差」を求める目的は，母集団のばらつきの大きさを知るためであった．しかし母集団のばらつきの大きさを完全に知るためには無限個のデータが必要であり，それは絶対に行えない．よって「標本分散」「標本標準偏差」を算出し，それを「母分散」「母標準偏差」の推定値とする．

ここで，母数を示す文字について説明する．これまでに説明してきたように「標本平均：$\bar{x}$」「標本分散：$s^2(x)$」「標本標準偏差：$s(x)$」のように統計量を示すためにはアルファベットが使われるが，それに対して母数に関してはギリシャ文字が用いられる．例えば「母平均：μ」「母分散：σ^2」「母標準偏差：σ」である．

ここで我々は測定を繰返し，その標本平均を測定結果とするわけであるが，ここで求めた標準偏差はあくまでも測定データが平均的にどのくらいばらつくかを示した指標である．しかし，測定結果を標本平均としているのであれば，標本平均がどの程度ばらつくのかを知る必要がある．標本平均がばらつく，というのは例えばある測定を 5 回繰返して標本平均を得たとしよう．そしてさらに 5 回繰返してまた標本平均を得る，これを繰返すといくつかの標本平均が得られる．そうすればいくつか得られた標本平均間のばらつきが求められるだろう．このばらつきを標本平均の標準偏差という．ただ，上記のように繰返し測定を何回も行って標本平均を幾つも得てから標本平均の標本標準偏差を求めるという手続きは非常に煩雑である．よって通常はこのようなことは行わず，次式によって求める．

$$s(\bar{x})=\frac{s(x)}{\sqrt{n}} \quad \cdots\cdots (6)$$

つまり，データの標準偏差を繰返し回数の平方根で割ることによって求められる．これは繰返し回数が多くなればなるほど標本平均のばらつきが小さくなることを意味しているが，例えばサイコロを振って出た目の平均値の標準偏差を求めることを考えると，3 回振ったときの平均値より 10 回振ったときの平均値のほうがばらつきは小さくなるだろう．なぜなら，もしサイコロを 1 回だけしか振らない場合は，1・6 という端の値が 1/6 の確率で出るが，3 回の平均値であれば 3 回連続 1・6，10 回の平均値であれば 10 回連続 1・6 が出なければ平均値が 1・6 にはならない．つまり端の値が出にくくなるということは，繰返し数が多くなるにつれて標本平均のばらつきが減り，その減る程度が $\sqrt{n}$ 分の 1 である，ということを意味している．

4　確率分布について

母集団のばらつきの推定値を求めることができたが，母平均，母標準偏差が分かれば母集団のことをすべて知ることができたと考えるのは早計である．なぜなら，母標準偏差の値が同じであっても，母集団の形は異なっているかもしれないからである．ここで母集団の形について言及しているが，その母集団の形を表すものを「確率分布」と呼ぶ．

「確率分布」は確率変数の値の取りやすさを表したものであり，典型的なものとして「矩形分布」「三角分布」などがある．図 1 に矩形分布，図 2 に三角分布を示す．

図 1 は母平均が 0，分布の半幅が 1 である矩形分布であり，これは $-1<x<1$ の範囲にすべての測定データが存在し，さらにその範囲内の値は同じ確率で出現することを意味している．

図 2 は母平均が 0，分布の半幅が 1 である三角分布であり，これは $-1<x<1$ の範囲にすべての測定データが存在することは先ほどの矩形分布と同様であるが，母平

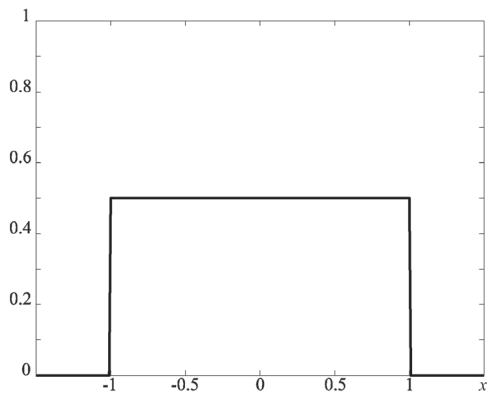


図 1　矩形分布

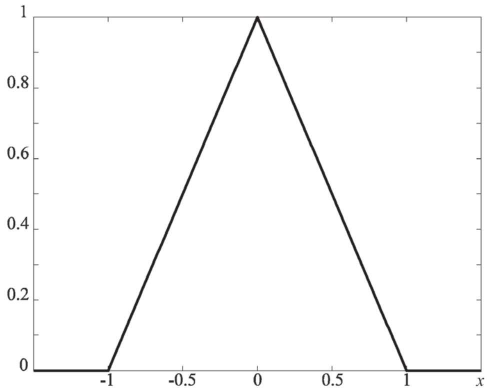


図 2　三角分布

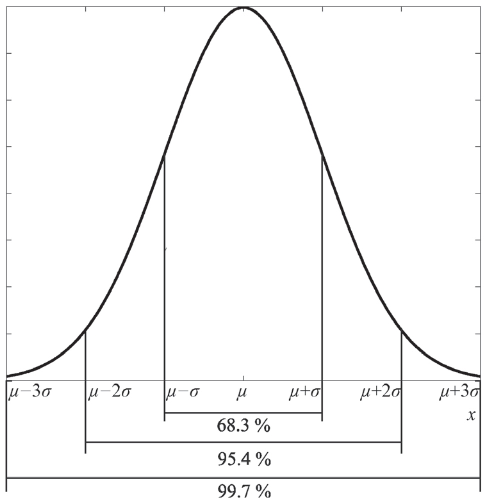


図 3　正規分布

均に近づくほど値の出現確率は高くなり，端に行くほど確率が低くなる分布である．

さて，ここで示された矩形分布も三角分布も両者とも値の存在範囲は，$-1 < x < 1$ である．しかし分布の形が異なるため，母標準偏差の値は異なる．矩形分布であれば分布の半幅を $\sqrt{3}$，三角分布であれば分布の半幅を $\sqrt{6}$ で割れば母標準偏差が求められる．つまり，図 1 の矩形分布の母標準偏差は $1/\sqrt{3}$，図 2 の三角分布の母標準偏差は $1/\sqrt{6}$ となる．よって母平均，母標準偏差が分かったとしても母集団について十分な情報を知ることになったわけではなく，確率分布の形を知らなければ存在範囲を知ることはできない．

それでは測定において最も重要な確率分布は何だろうか？それは正規分布である．正規分布を図 3 に示す．正規分布は多くの測定結果は正規分布に従っていることが知られているように，非常に汎用性のある分布であり，測定結果がどの程度信用できるかを求めるために用いられる最も基本的な分布である．正規分布は図 3 で示したようにきれいな対称形の山形の分布である．

正規分布は母平均と母標準偏差（母分散）が分かれば一意に決定する分布であり，図 4 に示すように，（母平均±母標準偏差）の範囲に約 68.3 %，（母平均±2×母標準偏差）の範囲に約 95.4 %，（母平均±3×母標準偏差）の範囲に約 99.7 % の値が存在することが知られている．基本的な母平均の存在範囲の推定は上記の性質を用いて行う．また 68.3 %，95.4 %，99.7 % 以外の確率を用いたい場合でも，母標準偏差を何倍したときにどの程度の確率が含まれるかを数表やソフトウェアを用いれば簡単に知ることができる．

5　正規分布を用いた母平均の区間推定

それではここから先ほど説明した正規分布を用いて，母平均の存在範囲を推定する方法を解説する．まず測定の母集団が母平均 μ，母標準偏差 σ の正規分布に従っていたとしよう．その測定の母集団を図 4 に示す．

図 4 で示した測定の母集団からデータを n 個サンプリングし，その標本平均を求めると，式（6）で示したように標本平均の標本標準偏差は $\sqrt{n}$ 分の 1 となるが，もちろん母標準偏差も同様で，

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}} \quad \cdots\cdots (7)$$

ここに，$\sigma(\bar{x})$ は標本平均の母標準偏差を表す．

が成立する．よって，図 4 の測定の母集団から n 個サンプリングして標本平均を求めたとき，その標本平均

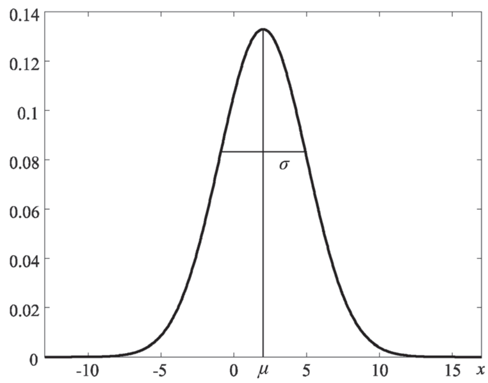


図 4　測定の母集団

上記の図では $\mu=2$，$\sigma=3$ としている．

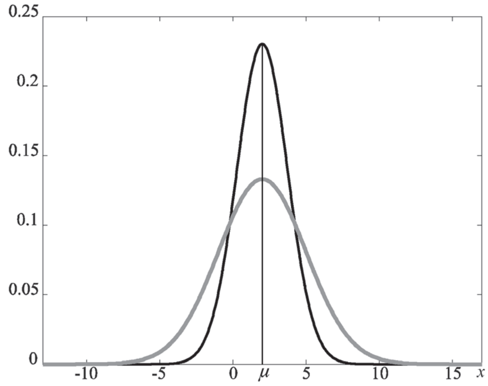


図 5　標本平均 $\bar{x}$ の確率分布

上記の図では $n=3$ としている．

の確率分布は図5になる．

図5でグレーの線が図4で示したデータの確率分布である．そこから n 個サンプリングし標本平均を求めているので，その分ばらつきが小さくなっている．

さてここで，標本平均の分布は母平均 μ，母標準偏差 $\sigma/\sqrt{n}$ を持つ正規分布に従っているが，このままでは測定それぞれによって母平均，母標準偏差の値が異なり，一貫性をもって演算を行うことが難しい．よってこの分布をさらにシンプルな分布へ変形し，その結果から演算を行うことを考える．ここで図5の標本平均の分布を確認すると，分布の中心は0からずれているので，中心を0にすることを考える．分布の中心は μ なので，$\bar{x}-\mu$ の分布を考えれば，分布の中心が0となるだろう．図6に $\bar{x}-\mu$ の確率分布を示す．

次に $\bar{x}-\mu$ の確率分布は母標準偏差が $\sigma/\sqrt{n}$ であるので，この母標準偏差を1にすることを考える．つまり，$\bar{x}-\mu$ を母標準偏差で割った $(\bar{x}-\mu)/(\sigma/\sqrt{n})$ の分布を考えればよい．図7に $(\bar{x}-\mu)/(\sigma/\sqrt{n})$ の確率分布を示す．

図7で示した確率分布は必ず母平均 $\mu=0$，母標準偏差 $\sigma=1$ の正規分布となる．この母平均 $\mu=0$，母標準偏差 $\sigma=1$ の正規分布のことを「標準（規準）正規分布」と呼び，今行ったような標準正規分布へとデータを変換することを「標準化（規準化）」という．正規分布の数表等はこの標準正規分布の性質を表したものである．

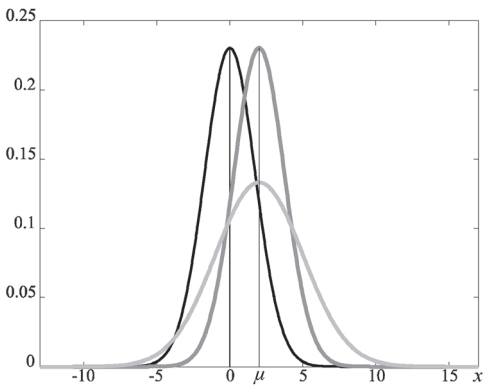


図6：$\bar{x}-\mu$ の確率分布

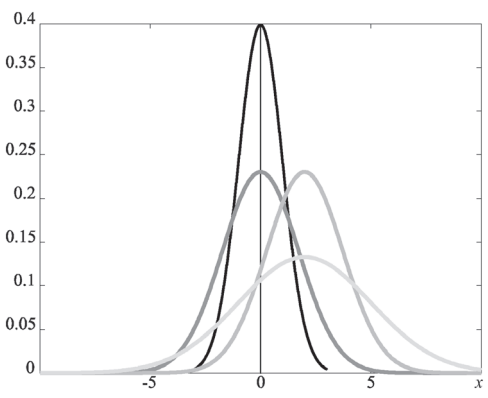


図7　$(\bar{x}-\mu)/(\sigma/\sqrt{n})$ の確率分布

次に標準正規分布と図3で示した正規分布の性質を考え合わせると，

$$-1<\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}<1 \quad \cdots\cdots (8)$$

の範囲内には全データの68.3 %，

$$-2<\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}<2 \quad \cdots\cdots (9)$$

の範囲内には全データの95.4 %，

$$-3<\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}<3 \quad \cdots\cdots (10)$$

の範囲内には全データの99.7 % が含まれる，ということが分かるだろう．この中で代表として，式（9）について考える．

式（9）を変形する．

$$-2<\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}<2$$

$$-2\frac{\sigma}{\sqrt{n}}<\bar{x}-\mu<2\frac{\sigma}{\sqrt{n}}$$

$$-\bar{x}-2\frac{\sigma}{\sqrt{n}}<-\mu<-\bar{x}+2\frac{\sigma}{\sqrt{n}}$$

$$\bar{x}-2\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+2\frac{\sigma}{\sqrt{n}} \quad \cdots\cdots (11)$$

式（11）を見ると，母平均 μ は $\bar{x}-2\sigma/\sqrt{n}$ から，$\bar{x}+2\sigma/\sqrt{n}$ までの範囲の中で，約95.4 % の確率で存在する，ということを意味していることが分かる．これが正規分布を用いた母平均の区間推定と呼ばれるものである．95.4 % 以外の確率を用いたい場合は，正規分布表やソフトウェアで簡単に知ることができる．

さて，ここまで説明してきた母平均の区間推定であるが，実際の測定データに適用することを考えると大きな問題が2点存在する．

まず1点目は，母平均の区間推定を行うためには，式（11）を計算しなければならない．そこで式（11）に含まれる変数を確認すると，$\bar{x}$ については，標本平均なので，実際に測定したデータの標本平均を求めればよいので問題ない．次に n については，これは標本平均を求めるための繰返し回数であるので，これも問題ない．最後に σ であるが，これは母数である．つまり無限回測定を行わない限り知ることができない．よって本来 σ は未知の値である．そうなると区間推定は行えなくなってしまい，この手法は意味のないもののように思えるが，そういうわけではない．つまり，この正規分布を用いた区間推定は，「非常に質のいい母標準偏差の推定値を事前に知っている」というときに用いることができる．例えば，ある製造ラインで長年製品を作製しており，そのラインで生産された製品のばらつきはこれまでに多くの製品を作製しているので，そのこれまで生産された製品の測定結果から求められた標本標準偏差は質の良い母標準偏差の推定値であることが担保されている，

というようなときである．これであれば，正規分布を用いた母平均の区間推定を行っても十分意味のある結果を得ることができる．言い換えると，顧客から預かった測定対象物について5回繰返し測定を行った，というようなときには使うことはできない．なぜなら5個のデータで母標準偏差を推定することになるが，その推定はデータの個数が少なすぎ，あまりうまくいかないことが考えられるからである．このようなときには t–分布を用いた母平均の区間推定を行う．これに関しては後述する．

もう一つの問題は，測定の母集団が正規分布に従っていない場合はどうするのか？　というものである．確かに今回はまず図4で示した正規分布に従った測定の母集団が存在し，そこから n 個データを取得して標本平均を算出したときの母平均の区間推定であった．そうであれば，測定の母集団が正規分布に従っていなければここで論じた母平均の区間推定に関しては全く使えなくなってしまう気がするが，実はそのようなことはない．それは「中心極限定理」が存在しているからである．

中心極限定理とは簡単に言うと，「一般的に，測定の母集団がどのようなものであっても，そこからいくつかのデータをサンプリングし，標本平均を求めた場合，標本平均の確率分布は，標本平均を求めるためのサンプリング数が多くなるにつれて，正規分布に近づいていく．」というものである．例を見てみよう．

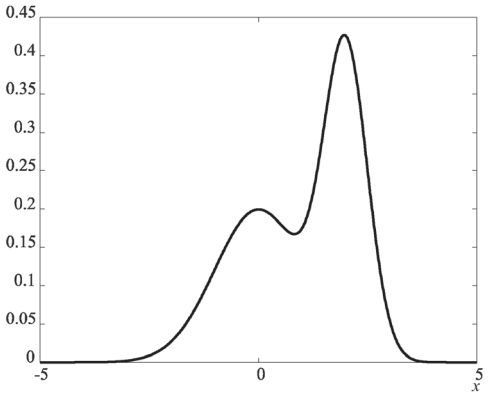


図8　ある測定データの確率分布

図8にある測定データの確率分布を示す．この確率分布は見て分かるように二つの山があり，正規分布から大きくかけ離れたような分布になっていることが分かるだろう．

この母集団から一つデータをランダムにサンプリングする，という手続きを1 000 000回繰返したときのヒストグラムと，その分布の母平均，母標準偏差を持つ正規分布を重ねて表記したグラフを図9に示す．

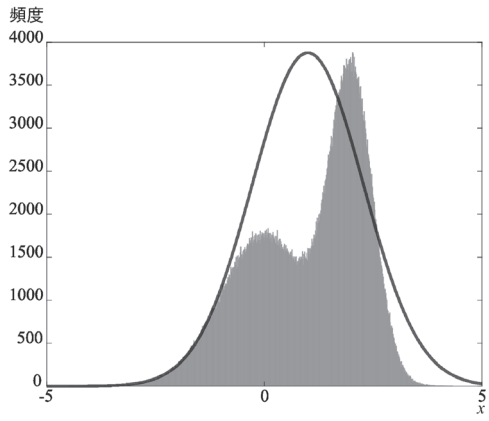


図9　サンプリング1回のときのヒストグラム
正規分布の縦軸はヒストグラムにスケーリング済み．

サンプリング数が1回なので当然ヒストグラムは元の確率分布と同じ形になる．また同じ母平均，母標準偏差を持つ正規分布とも全く形が異なることが分かる．

次に測定データの確率分布から2個，3個，4個，5個サンプリングし，それぞれ平均値を求める，という手続きを1 000 000回繰返したときのヒストグラムと，その分布の母平均，母標準偏差を持つ正規分布を重ねて表記したグラフを図10（a）～（d）に示す．

これを見ると，繰返し数が2回のときはまだ山が2つ見えているが，繰返し数が3回になれば山が一つになり，ほぼ同じ母平均，母標準偏差を持つ正規分布と変わらない分布になっていることが分かる．さらに繰返し数が4回，5回と増えていくにつれて正規分布との一致度が高くなる．繰返し数が5回のときはほぼ正規分布と区別がつかなくなっていることが分かるだろう．

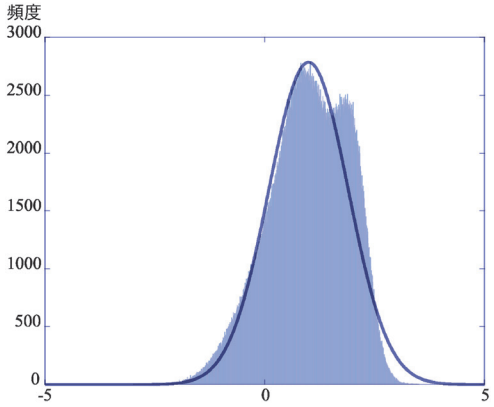
このように中心極限定理は図8で示したような変わった分布であっても強力に働き，繰返し数が3，4回くらいになるとほとんど正規分布とみなしても問題ないくらいになる．この中心極限定理の成立条件は，測定データの分布の分散（標準偏差）の値が有限，というものである．通常の分布であれば，ばらつきの値が有限であることは当然であるので成立条件を満たす[*2]．

このように，通常の測定結果であれば，中心極限定理の成立条件は満たしていると考えられ，標本平均は正規分布に従っていると考えても差し支えない．

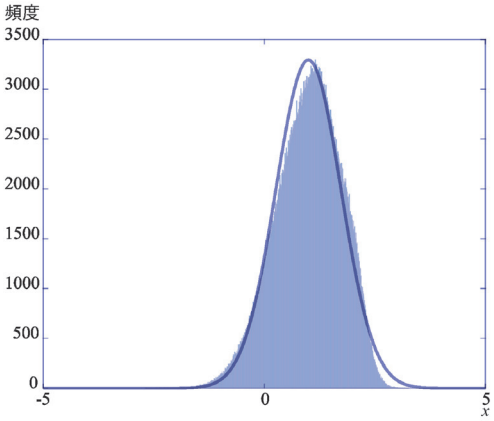
6　t–分布を用いた母平均の区間推定

さて先ほど説明したように，正規分布を用いた母平均の区間推定は十分質の良い母標準偏差の推定値を知っているということが前提になっていた．しかし，このような前提条件はいつもクリアできるようなものではない．例えば，顧客から持ち込まれたサンプルを測定する場合は，多くの繰返しを行うことは難しいだろう．そのよう

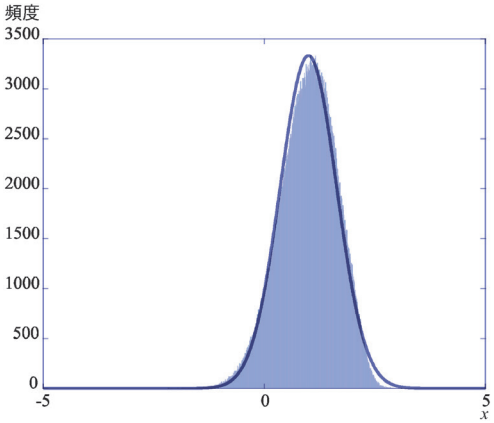
[*2] 中心極限定理の成立条件を満たさない例外的な分布の例としては，後述する自由度が1のときの t- 分布が相当し，特にこの分布のことをコーシー分布と呼ぶ．コーシー分布は分散が存在しないだけでなく，平均値（厳密には期待値）も存在しない．



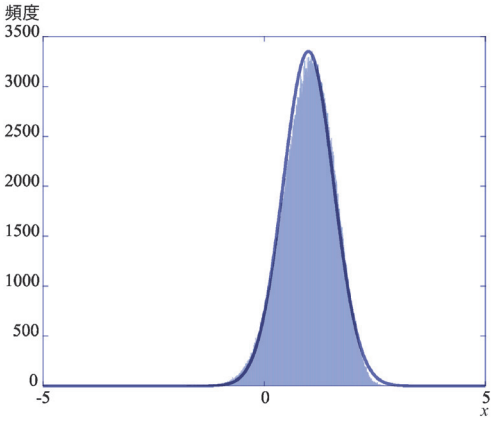
（a）繰返し 2 回



（b）繰返し 3 回



（c）繰返し 4 回



（d）繰返し 5 回

図 10　標本平均のヒストグラム
正規分布の縦軸はヒストグラムにスケーリング済み．

な場合は，標本標準偏差を求めることはできても，その標本標準偏差を母標準偏差の質の良い推定値と考えるのは無理がある．そのような場合はどうすればよいのだろうか？

それは，先ほどの正規分布を用いた推定の際は，母標準偏差を用いて正規化し，母平均 0，母標準偏差 1 の正規分布に変形したが，今回の場合は，標本標準偏差を用いて正規化を行えばよい．つまり，

$$\frac{\overline{x}-\mu}{s(x)/\sqrt{n}} \quad \cdots\cdots (12)$$

の値を求めればよい．

正規分布のときは，正規化した値は標準正規分布に従っていたが，式（12）にて正規化した値は，自由度 $n-1$（繰返し回数から 1 を引いた，標本分散，標本標準偏差を算出する際に用いる自由度と同じもの）の t–分布に従うということが知られている．t–分布を図示したものを図 11 に示す．

図 11 に表記されている線は，最も色の薄いものは自由度が 3，色が濃くなるにつれて自由度が大きくなり，最も色の濃いものが自由度無限大（正規分布）のものである．つまり，自由度が小さい場合は正規分布と比べ幅が広い，つまり算出された標本標準偏差にもばらつきが存在し，それが反映されて分布の幅が広くなっている．そして自由度が大きくなるにつれて正規分布に近づいていく．最後に自由度が無限大になるとデータをすべて知っているという状態になるため，そうなると自由度が無限大の t–分布と正規分布が一致するのは当然である．

t–分布は先ほどの正規分布とは違い，自由度が必要となる．自由度が決定すれば一意に t–分布は決定する．

t–分布を母平均の区間推定に用いるためには次式を用いる．

$$-t(p,\nu)<\frac{\overline{x}-\mu}{s(x)/\sqrt{n}}<t(p,\nu) \quad \cdots\cdots (13)$$

式（13）を正規分布のときと同様に式を変形すると，

$$\overline{x}-t(p,\nu)\frac{s(x)}{\sqrt{n}}<\mu<\overline{x}+t(p,\nu)\frac{s(x)}{\sqrt{n}} \quad \cdots\cdots (14)$$

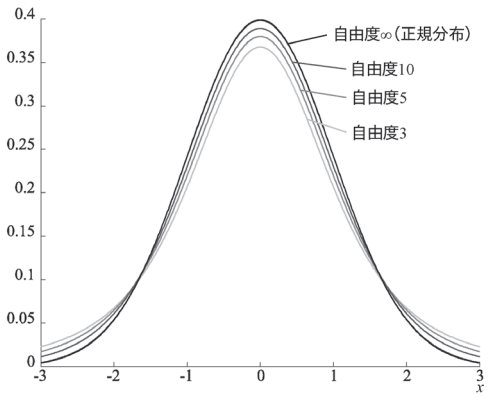


図 11　t–分布
自由度 $\nu=3, 5, 10, \infty$［正規分布］．

表 1　t–分布表

自由度：ν	確率：p		
	0.95	0.99	0.999
1	12.706	63.657	636.619
2	4.303	9.925	31.599
3	3.182	5.841	12.924
4	2.776	4.604	8.610
5	2.571	4.032	6.869
6	2.447	3.707	5.959
7	2.365	3.499	5.408
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
20	2.086	2.845	3.850
50	2.009	2.678	3.496
100	1.984	2.626	3.390

となる．つまり，母平均が含まれる確率 p と $s(x)$ を算出したときの自由度 ν を求め，そこから $t(p, \nu)$ の値を求めればよい．$t(p, \nu)$ の値は数表やソフトウェアで簡単に求められる．一例を表 1 に示す．

例えば，データを 10 個取得し，標本平均と標本標準偏差を算出した場合，95 % の確率で母平均が含まれる区間を求めたいとすると，この場合自由度は 9 となるため，表 1 の「自由度：9」「確率：0.95」の欄にある，$t(0.95, 9) = 2.262$ を採用して式（14）に代入すれば母平均の区間推定を行うことができる．

7　最　後　に

ここまで駆け足で統計の基礎から母平均の区間推定まで説明したが，最初に述べたように重要な点も大きく省いて説明を行っている．例えば，母平均，母分散にかかわる期待値の議論や，確率分布にかかわる確率密度関数，累積分布関数等の議論を行っていない（例えば，本稿での確率分布を表したグラフの縦軸については全く触れていない）．もっと詳細を知りたい方は統計の教科書等で勉強していただきたい．とりあえず本稿の内容が分かれば，初級程度の統計の教科書は読むことができるだろう．

統計はデータを解釈するうえで非常に重要なものである．しかし測定データを取得し，その後そのデータを解釈しようとしたときにはじめて統計をどうしようか？と考える人がいる．しかしそれは是非やめてほしい．なぜなら，測定データは統計的手法に合わせて取得法を決めることができるからである．つまり，データを取得する前にデータ処理方法を決めておけば，そのデータ処理方法を適用しやすい測定データを取得する方法を考えることができる．そのようなデータであれば非常に単純な統計的手法を適用するだけでも質の良い結果を得ることもできる．

測定を行う前にはもちろんその測定の定義，測定方法，測定手順を決定し測定を行うだろうが，それとともに取得したデータにどのような統計的手法を適用するのか，ということも併せて決定し，そのあとに測定を行っていただきたい．

文　献

1）田中秀幸：“分析・測定データの統計処理”，(2014)，(朝倉書店)．

田中 秀幸（Tanaka Hideyuki）
産業技術総合研究所計量標準総合センター工学計測標準研究部門データサイエンス研究グループ（〒305−8563 茨城県つくば市梅園 1−1−1 中央事業所 3 群）．筑波大学大学院工学研究科修了．博士（工学）．《現在の研究テーマ》測定における不確かさ評価に関する研究．《主な著書》分析・測定データの統計処理．