データ評価のための統計的方法

――確率分布と平均値の推定・検定――

田　中　秀　幸

1　はじめに

前回は，統計的手法を適用するために意味のあるデータをどのように取得するのかについて，母集団と標本について，期待値・分散・標準偏差について解説した。

今回は，統計的推定・検定の基礎となる確率分布とその確率分布を用いた推定・検定について解説する。

2　確率分布

測定データを取得したとき，そのデータのばらつきを視覚的に表すために，横軸にデータの値，縦軸に頻度をプロットしたヒストグラムがよく用いられる。図1にヒストグラムの例を示す。

このヒストグラムは標本に対して作成されるものである。では，これを母集団を表すものとして考えるとどのようなものになるであろうか？　母集団に対してヒストグラムを作成しようとすると縦軸の頻度の部分が作成できない。なぜなら，母集団には無限個のデータが含まれるからである。よって，母集団について表すときには縦軸を頻度ではなく，確率にすればよい。簡単な例としてサイコロの出る目の確率分布を考えよう。サイコロの目は1～6までの目があり，それぞれ1/6の確率で現れる。また，1～6以外の目が出る確率は0である。これをグラフで表すと図2になる。

図2で表されたものはサイコロの目の確率分布であるが，サイコロの目は実際の測定結果とは大きく異なる性質がある。それはサイコロの目は1，2，…，6といった飛び飛びの値のみ現れ，間の値は現れない。このような分布のことを離散分布と呼ぶ。しかし，実際の測定結果は飛び飛びの値ではなく，間の値もすべて含まれる。このような分布を連続分布という。

統計的な推定を行うときは事前に測定結果がどのような確率分布に従っているかということを決定し，その確率分布の性質を用いて推定を行う。

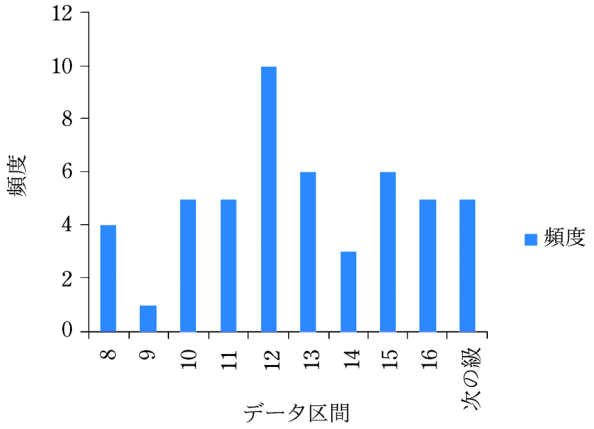

図1　ヒストグラムの例

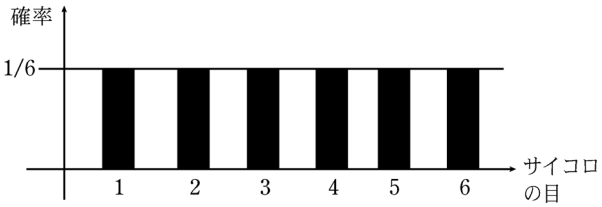

図2　サイコロの目の確率分布

3　正規分布

統計的推定はその測定結果がどのような確率分布に従うのか，ということを決定し行うが，その際にもっともよく用いられる確率分布が正規分布である。

正規分布とは，きれいな山形をした分布のことを表しており，その確率分布を式で表すと，

$$p(x)=\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} \quad \cdots\cdots\cdots\cdots\cdots\cdots(1)$$

となる。μは母平均，σは母標準偏差を表す。ここで$p(x)$は前回も簡単に説明した確率密度関数である。この確率密度関数をグラフで表したものが図3である。

式(1)を見て分かるように，正規分布は母平均μと母標準偏差σ（または母分散σ^2）が分かっていれば一意に決まる。つまり，測定結果が正規分布に従っているという前提条件があり，妥当な母平均，母標準偏差の推定

Fundamental Knowledge for Reliable Analysis—Statistical Methods for Data Analysis: Part 2.

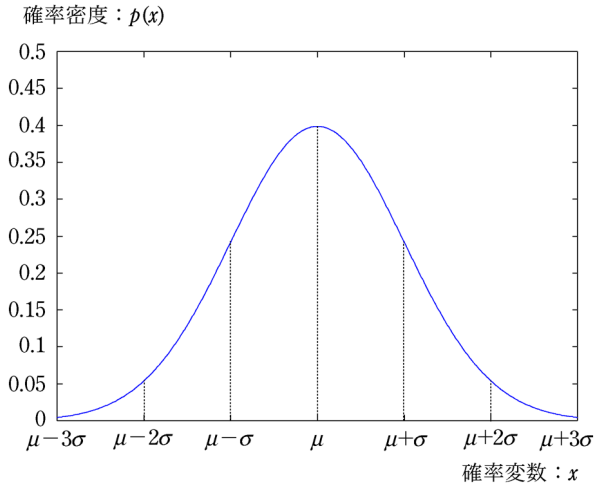

図 3　正規分布

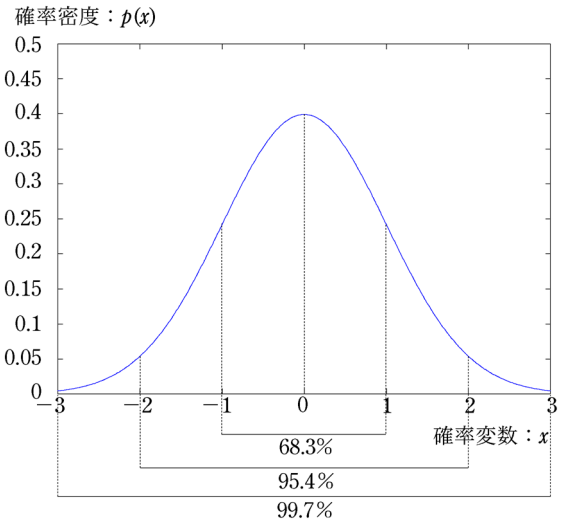

図 4　規準正規分布

値が分かっているのであれば，母集団の形を十分な精度で知ることができるということである。ここで言う妥当な母平均，母標準偏差の推定値は前回説明したように，測定量の定義，測定方法，測定手順を定め，それを実現した測定を十分な回数行えば知ることができる。次に測定結果が正規分布に従うのかどうかということであるが，これに関してはほとんどの測定結果は正規分布に従うと考えてもよい。ある測定を何回か行うとその測定結果のヒストグラムが平均値を頂点とした山形となることを経験している読者も多いであろう。またそれだけではなく，よく管理された測定における測定結果の分布が理論的に正規分布に近づくということが分かっている。つまり「よく管理された測定結果である」ということは測定にばらつきを与える要因が管理され，未知のコントロールできないばらつきを与える複数の要因が合成され，それが測定結果に影響を与えて繰り返しのばらつきとして現れるということを意味している。このようなときには中心極限定理という統計の定理が働く。これは同じくらいのばらつきの大きさを持つ複数の要因（この要因はどのような確率分布を持っていてもよい[*1]）の確率分布が合成されるとその合成された分布は正規分布に近づくというものである。よって，よく管理された測定結果の確率分布は正規分布に従うと考えても差し支えない場合が多い。

次に正規分布に関する計算を行うことを考えよう。正規分布に関する計算は式(1) を基にして行うのであるが，実際に式(1) を用いて計算することは少ない。

[*1] 実際にはどのような確率分布でもよいというわけではない。しかし，確率分布の母分散が有限であるというような条件はつくが，一般的な分布であれば問題なく中心極限定理が働く。例えば U 字分布と呼ばれるような分布の両端の確率が高く，分布の中心（母平均）の周辺が一番確率が低くなる分布であっても中心極限定理が働き，たとえ U 字分布のみをいくつか合成したとしても合成された分布は正規分布に近づく。

式(1) 内の確率変数 x を次のように変形する。

$$u = \frac{x-\mu}{\sigma} \quad \cdots\cdots(2)$$

式(2) を式(1) に代入すると，

$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-1/2\,u^2} \quad \cdots\cdots(3)$$

となる。式(3) は規準正規分布と呼ばれ，母平均 0，母標準偏差 1 の正規分布を表す。つまり，正規分布に従っている測定結果であれば，式(2) で表される変数変換（規準化）を施すと式(3) で表される規準正規分布で表すことができる。この規準正規分布を図 4 に示す。

図 4 に示されている 68.3%，95.4%，99.7% とは，規準正規分布に従う測定結果の中で ±1 以内に入るデータは全データ中の 68.3% ということを表している。他の確率も ±2 以内に 95.4%，±3 以内に 99.7% と同様である。規準正規分布の標準偏差は 1 であるので，±1，2，3 というのは標準偏差の ±1，2，3 倍ということと等しい。つまり，一般的な正規分布に関しても標準偏差の±1，2，3 倍の中にそれぞれ 68.3%，95.4%，99.7% 含まれる。この区間内に母平均が含まれる確率のことを包含確率という。これは前回にも提示した式

$$P_{a\le x\le b}(x) = \int_a^b p(x)\,dx \quad \cdots\cdots(4)$$

によってその他任意の包含確率の計算することができるが，規準正規分布は非常によく用いられる確率分布であるので，規準正規分布に関する包含確率が数表としてまとめられている。数値計算のソフトウェアでも規準正規分布に関する値を算出する関数がどのようなものでも用意されているはずである。規準正規分布でない場合も式(3) で示された規準化を行った後に計算を行えば規準正

規分布の数表がそのまま使える。よって，実際には式(1)を用いて計算を行わなければならない場合は非常に少ない。つまり，正規分布に従っているということであれば，その正規分布の母平均と母標準偏差だけが分かれば計算をするために十分である。よって，一般的には母平均 μ，母標準偏差 σ に従う正規分布のことを式(1)を用いて表すよりは，

$$\mathrm{N}(\mu, \sigma^2) \quad \cdots\cdots(5)$$

と表すほうが多い。例えば，規準正規分布は，$\mathrm{N}(0, 1^2)$ となる。

4　正規分布を用いた平均値の推定

先ほど解説した正規分布の性質を用いて母平均がどの区間にどのくらいの確率で含まれるのか，ということを推定することを考える。

測定結果 x_i は正規分布からのサンプリングであると考えても差し支えなく，その標本平均と標本分散がそれぞれ，

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \cdots\cdots(6)$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad \cdots\cdots(7)$$

であったとする。また，このときの標本分散は母分散の推定値と見なしても問題なかったとしよう。つまり，

$$\sigma^2 \approx s^2 \quad \cdots\cdots(8)$$

とする[*2]。

ここで標本平均の分布を考える。標本平均の母平均は μ と一致することは自明であろう。また，標本平均の母分散は前回解説したように，

$$\sigma^2(\bar{x}) = \frac{\sigma^2(x)}{n} \quad \cdots\cdots(9)$$

となる。また，測定結果が正規分布に従うのであれば，その測定結果の標本平均も正規分布に従い，その分布は $\mathrm{N}(\mu, \sigma^2(\bar{x}))$ となる。これに式(2)で表される規準化を行うと，

表 1　正規分布表

包含確率（%）	$\pm z$
80	1.282
90	1.645
95	1.960
99	2.576
99.8	3.090
99.9	3.291
99.98	3.719
99.99	3.891

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

となる。基準化された u は $\mathrm{N}(0, 1^2)$ に従う。次にどのくらいの推定精度で平均値の推定を行うかを決定する。ここでは 95.4% の確率で母平均が含まれる区間を考えてみよう。そうすると図 4 の規準正規分布の確率分布から，

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

となる。これを変形すると，

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

$$\bar{x} - 2\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2\frac{\sigma}{\sqrt{n}} \quad \cdots\cdots(12)$$

となる。式(12)は母平均が 95.4% の確率で含まれる区間を示している。これが正規分布を用いた平均値の区間推定である。

図 4 で示されている包含確率以外で区間推定を行いたい場合には正規分布表を用いればよい。正規分布表の例を表 1 に示す。ここにある包含確率に対応する z の値が標準偏差の何倍であるかに相当するものである。つまり，式(12)を一般式に書き換えると，

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}} \quad \cdots\cdots(13)$$

となる。

5　正規分布を用いた検定

正規分布を用いた検定とは，求められた標本平均は母平均と一致していると考えてもよいのかどうかということを判定する手法である。つまり標本平均はあくまでもサンプルから求められたものであり，母平均と完全に一致することはまずあり得ない。しかし，そこで求められた標本平均と母平均の差は，たまたま含まれるばらつきによるものであるのか，それとも，異なる母集団から取

[*2] 母分散は推定を行う際に求めた標本分散を母分散の推定値として用いる以外に，事前の測定結果により求められた母分散の推定値を用いてもよい。ただしその場合は，推定を行う際の測定データは事前の測定結果によって構成される母集団からのサンプリングでなければならない。つまり，事前に行った測定と，推定を行う際の測定の測定量の定義，測定方法，測定手順が異なっていてはいけない。

られたサンプルの平均値であるために値が異なるのかを判定するということである。つまり得られたサンプルは想定している母集団から取られたのかどうかということを判定するわけである。

想定している母集団を $N(\mu, \sigma^2)$ とする。ここでは母平均と母標準偏差は既知[*3]である。このとき検定を行いたい対象の測定結果を

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

とする。次にここで得られている値を用いて規準化を行う。

$$u_0 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \quad \cdots\cdots(14)$$

ここで求められた u_0 は図 4 の横軸の値を表している。よって，例えば 95% の確率（検定のときには「5% の危険率」という言い方をする）で母平均と標本平均が一致するかどうかを判定したい場合には表 1 の正規分布表より $z = \pm 1.960$ であるので，$z = \pm 1.960$ 内に u_0 の値が含まれるかどうかを判断すればよい。もし含まれるならば 5% の危険率で標本平均と標本分散が等しいと言える。また含まれないのであれば，等しいとは言えない。

この手法が製品生産の品質管理法の基礎となっている。

6　t 分布を用いた推定・検定

正規分布を用いた推定・検定を行う際に前提条件となるのが，測定結果が正規分布からのサンプリングであると考えられるかということと，その正規分布の母標準偏差（検定ではさらに母平均）が既知であるということである。先ほど説明したように，良く管理された測定における測定結果は正規分布をしていると見なしてもかまわないということであったが，母標準偏差が既知であるという条件はどうだろうか？　こちらに関しては母標準偏差が既知であるという条件が満たせない場合も多く存在する。

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どうだろうか。顧客から持ち込まれたサンプルは分銅等の手軽に繰り返し測定ができるものであればよいが，土壌サンプルなどの場合はサンプルをいくつか用意してもらったとしても，母標準偏差を推定するに当たって十分な数が用意されているとは限らない。よって，測定結果は正規分布に従っているということは分かっているが，サンプル数が少なすぎるがために標本標準偏差が母標準偏差の推定値であると見なすことができないという場合がある。このような場合は t 分布を用いた推定・検定を行う。t 分布とは先ほどの正規分布と非常によく似た分布であるが，規準化を行うとき正規分布では母標準偏差 σ を用いるが t 分布では標本標準偏差 s を用いる。つまり，

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad \cdots\cdots(15)$$

によって規準化を行うということである。このように規準化された t の値は t 分布をするということが分かっており，その確率分布は図 5 のようになる。

図 5 には複数の t 分布の確率分布が描かれているが，これはそれぞれ自由度が異なっている。自由度とは標本標準偏差を算出するためのデータ数から 1 を引いた値である（前回参照）。つまり，データ数が少ない場合には情報の量が少ないので幅が広い分布となるが，データ数が多くなるにつれて情報が増え，幅が狭い分布となる。さらにデータ数が無限大になったときには t 分布は規準正規分布を完全に一致する。これは当然のことであろう。自由度無限大のデータから算出された標本分散は母分散と等しいからである。

推定・検定の方法は正規分布の手法と全く同じであるが，表 1 にあげた正規分布表を用いるのではなく，表 2 に示す t 分布表を用いる必要がある。ただし，t 分布表は先ほどの正規分布表とは異なり，自由度によって値が異なるので注意してほしい。

t 分布による推定は，測定結果 $\bar{x}$ とその標本標準偏差

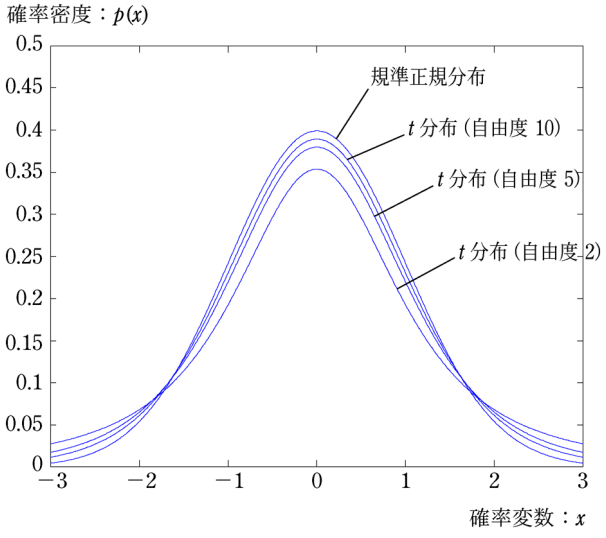

図 5　t 分布

[*3] 本来であれば母平均と母標準偏差が完全に分かるということはあり得ない。しかし十分な回数の事前測定が行われており，その平均値と標準偏差を母平均と母標準偏差と見なしても差し支えないという状況はよくある。例えばある工場で生産している製品について，日々製品を測定し，その結果から母平均と母標準偏差を推定しているという場合などである。このとき，ある日作成された製品が正常であるか（その日の製品の標本平均と母平均が等しいと見なせるか）をチェックするということがこの検定で行える。

表 2　t 分布表

自由度	包含確率（%）				
	90	95	99	99.9	99.99
1	6.31	12.71	63.66	636.62	6366.20
2	2.92	4.30	9.92	31.60	99.99
3	2.35	3.18	5.84	12.92	28.00
4	2.13	2.78	4.60	8.61	15.54
5	2.02	2.57	4.03	6.87	11.18
6	1.94	2.45	3.71	5.96	9.08
7	1.89	2.36	3.50	5.41	7.88
8	1.86	2.31	3.36	5.04	7.12
9	1.83	2.26	3.25	4.78	6.59
10	1.81	2.23	3.17	4.59	6.21
20	1.72	2.09	2.85	3.85	4.84
30	1.70	2.04	2.75	3.65	4.48
40	1.68	2.02	2.70	3.55	4.32
50	1.68	2.01	2.68	3.50	4.23
100	1.66	1.98	2.63	3.39	4.05

s，母平均 μ から，

$$\bar{x}-t\frac{s}{\sqrt{n}}<\mu<\bar{x}+t\frac{s}{\sqrt{n}} \quad \cdots\cdots(16)$$

によって行い，検定は

$$t_0=\frac{\bar{x}-\mu}{s/\sqrt{n}} \quad \cdots\cdots(17)$$

を算出し，t_0 と表 2 の t の値と比較することによって行う。

7　最 後 に

今回は，第 1 回の統計の基礎をベースとして確率分布，正規分布・t 分布を用いた推定と検定について解説した。正規分布は統計における一番基礎的な分布であり，正規分布を基にした様々な手法が開発されている。もちろん t 分布もその一つである。

今回解説した推定と検定であるが，他にもいろいろある推定・検定法も考え方はほぼ同様である。つまり確率分布を設定し，その確率分布が正しいとしたらこのような値が出るはずである，という基本的な考え方は変わらない。よって，いろいろな手法を学ぶ前にまずこの正規分布に関する推定・検定をしっかり自分のものにしてから他の手法の勉強を行ってほしい。

次回は，化学分析におけるデータ解析で非常に重要な手法である分散分析法の基礎と，分散分析法を用いた標準物質への値付け法について解説したいと思う。

田中秀幸（Hideyuki Tanaka）
産業技術総合研究所計測標準研究部門物性統計科応用統計研究室（〒305-8563 茨城県つくば市梅園 1-1-1　産総研中央第 3）。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≫計測における不確かさについて。

新刊紹介

専門基礎：化学入門
――その論理と表現――

藤原鎭男 著

本書は，化学を専攻・学習するために大学入学を志す高校生，化学以外の専門教育を受けた後，これから化学専攻の大学院で学ぼうとする人たちへの道しるべとして書かれたものである。化学の成長は，「元素の周期律」を生んだ「帰納的手法」と，「原子構造」を生んだ「解析（演繹）的手法」との協調によってなされた，とする著者の主張が全巻を貫いている。具体的な内容は以下のとおりである。「第 1 部　化学の論理」は，はしがき，元素の周期律，原子構造の 3 章から構成されている。「第 2 部　近代科学の基本量」では，科学の知識の体系化を図る際の土台ともいえる基本量を取り上げている。第 3 部では，科学知識の表現に重要な役割を果たす文字（文章），数値，画像の三媒体について述べられている。補遺の前半では，「化学は実証科学であり，実験が基盤である。実験を無視した化学はありえない。」との考えに基づき，実験への取り組み方が詳述されている。次いで後半では，科学を何故学ぶか，どう学ぶか，についての助言が述べられている。化学の入門書としてのみならず，科学（学問）の入門書として一読をお勧めしたい著書である。

(ISBN 978-4-567-20170-4・A5 判・115 ページ・1,800 円＋税・2008 年刊・廣川書店)