―基礎科学から医学・医療を見る―

正規母集団，推定の考え方

儀我真理子
日本医科大学基礎科学数学

Normal Population, Estimations

Mariko Giga
Department of Mathematics, Nippon Medical School

Abstract

The main part of the statistics is an investigating of the population by using samples. In this issue we state about a normal population first. In statistics we assume that a population follows a normal distribution in most cases. It is rational practically and theoretically. Another subject is the idea and methods of estimation (point estimation and interval estimation). We understand that point estimation itself is seldom used in the medical scene. However, the maximum likelihood estimation that is considered under the claims in point estimations is used as a value in several statistical methods. In interval estimations we can discuss the accuracy also.
（日本医科大学医学会雑誌　2014; 10: 16–20）

Key words: normal population, central limit theorem, point estimations, maximum likelihood estimation, interval estimations

1. 初めに

ここでは，統計の大きな分野の一つである“推定”の基礎的なこと，および統計処理において多くの場合仮定される正規母集団について述べる．

推定には，点推定と区間推定がある．点推定はサンプルから１つの値例えば母平均などの推定値を求めることで，母集団の分布の仮定は必要ない．点推定はごく大雑把なものを見るのに分かりやすい．ただ精度がわからないという危うさがあるので，きちんとした医学の話には推定自身としてはあまり使われないようである．しかし様々な統計手法を扱う過程において点推定の考え方が使われることは多い．区間推定においては母集団の分布の仮定が必要である．それだけの仮定をおいているので，区間推定では推定の精度も合わせて表現できる．

一般に統計理論においては，母集団分布は正規分布に従っていることを仮定する．実際世の中のたくさんのものが多くのデータをとると正規分布に近づくことが見てとれるが，そのことは理論的にはっきり裏付けられている．

母集団分布が正規分布に従うことが仮定できない場合や，サンプル数が少なく正規母集団に従っていると仮定することが難しい場合には，ノンパラメトリック検定が用いられる．

文字の使い方であるが，全体において，確率変数は大文字で，その実現値は小文字で書いている．

Correspondence to Mariko Giga, Department of Mathematics, Nippon Medical School, 2–297–2 Kosugi-cho, Nakahara-ku, Kawasaki, Kanagawa 211–0063, Japan
E-mail: mariko@nms.ac.jp
Journal Website（http://www.nms.ac.jp/jmanms/）

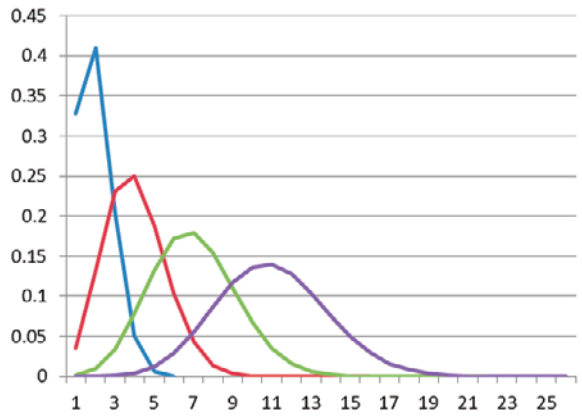

図 1

2. 正規母集団

私たちは，例えば，身長は高い低い，血清アルブミン濃度は濃度の値が大きい小さい，というように，2つの対立概念の兼ね合いで量の大小を表現する．だからそれらの各々の確率の分布を離散的に表した 2 項分布 $B(n, p)$ はすべての基本となる自然な分布と言える．ここで $B(n, p)$ の平均は np，分散は $np(1-p)$ である．個数 n を大きくするとその極限はある分布すなわち正規分布に近づく．詳しく言うと次の定理の形になる．

定理 2　ラプラス（Laplace）の定理

確率変数 $X_1, X_2, \cdots$ が 2 項分布 $B(n, p)$ に従うとき，

$$\lim_{\kappa\to\infty} P\left(a \le \frac{X_\kappa - np}{\sqrt{npq}} \le b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

上の**図 1** は 2 項分布 $B(n, 0.2)$ の $n=5$，15，30，50 としたときのグラフである．$p=0.2$ なので一番左側のもの（$n=5$）はかなり偏ったグラフになるが，n を大きくしていくと左右対称なグラフに近づく．自然な分布である 2 項分布の極限なので，正規分布も自然な分布と言えるはずである．また正規分布は誤差分布とも言われ，すべてのことに必ずおこってしまう誤差を表現しているととらえられる．

正規分布に関しては，次の重要な定理が成り立つ．

定理 3　中心極限定理（central limit theorem）

X_1，$X_2, \cdots$ は，互いに独立で同じ分布をもつ確率変数列であるとする．その平均値を μ，分散を σ^2 とおくとき，分散が有限ならば，$a<b$ なる任意の a，b に対して，次が成り立つ．

$$\lim_{n\to\infty} P\left(a < \frac{\frac{X_1+\cdots+X_n}{n} - \mu}{\sqrt{\frac{\sigma^2}{n}}} < b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

（左辺の括弧の中の形については，後に出てくる命題 5 を参照のこと．）

中心極限定理は一口で言うと“多数の，互いに独立な確率変数の平均を作ると，その分布は正規分布になる”という驚くべき定理である．元の確率変数の分布には何の仮定もおいていないのに，個数を増やしていくとそれらの平均は正規分布という特別な分布に近づくという結論が得られるのである．

中心極限定理はこの形で述べられることが多いし，母集団から取り出した n 個の標本を確率変数 X_i（$i=1, \cdots n$）と見て総計理論を作るときには，各 X_i の分布として母集団の分布を考えるのでこの形がよいであろう．でも実は仮定を少し弱めることができる：$X_1, X_2, \cdots$ は同じ分布でなくても一様有界ならばよい．すなわち起こりうるすべての X_i において $|X_i|$ が共通のある数で抑えられていればよい．

例えば，いくつもの科目の平均点の分布を考えると，正規分布に従うとよくいわれる．これは中心極限定理より正しい．しかし分布が同じであるという仮定をつけることは，このような現実の問題では応用しにくいことも多い．その点そこまで要求しなくていいのなら使いやすいと思う．

ところで，いくつかの科目の平均点の分布以前にひとつの科目の点数の分布も正規分布に従うことが多い．これも中心極限定理で説明がつくであろうか．テストにおいては，本人の努力，勉強時間，勉強環境，勉強の仕方，教師の教え方，持って生まれた素質などたくさんの要素が合わさって点数という結果が得られると考えることができる．だからそれらを合わせた意味で“中心極限定理からひとつの科目の点数の分布も正規分布になる”と考えられる．どんなことでも，それが起こるにはたくさんの要因がある．この考え方が，要素の個数が十分多ければその集団は正規分布と見なせることの理由と言っていいであろう．

標本数を増やすと左右対称な正規分布に近づくことが多いが，中には右側に裾を伸ばした形になるものもある．その場合，対数正規分布に従っていると仮定したほうがいいときがある．これは確率変数の自然対数をとると正規分布に従うものである．対数正規分布に

従うか否かを見る簡便な方法としては，対数正規確率紙にプロットして直線になるかを見るやり方がある（[5, p40–45]，[8, p37–42] 参照）．医学で扱われるものには指数関数的なものが多いので，散らばり（誤差）の大きさも考えている量に比例することが比較的多いと言える．例えば，血清タンパクやクレアチニンの血液生化学検査値などは対数正規分布に従うと考えられる．

3. 点推定

母集団の様子を表す数値として，平均，分散などがあるが，これらの値を母数 θ と呼ぶ．標本から母数の値を予想する．標本から求める母数 θ を推定する統計量 $\Theta = \hat{\theta}(X_1, \cdots, X_n)$ を定め，その値 $\hat{\theta} = \hat{\theta}(x_1, \cdots, x_n)$ によって θ を推定することを点推定という．

点推定では母数を一番好ましい値で推定したいわけだが，何をもって“一番好ましい”と言うかには，いくつかの考え方がある．ここでは代表的な次の［1］［2］［3］を挙げる．

［1］不偏推定量であること

その平均が母数 θ に一致する統計量 Θ を，不偏推定量（unbiased estimator）という．すなわち，Θ が θ の不偏推定量であるとは，

$$\mathrm{E}(\Theta) = \theta \tag{1}$$

を満たすことである．ここで E は平均を表す．

標本平均 $\bar{X}$ は母平均 μ の不偏推定量であるが，標本分散 S^2 は母分散 σ^2 の不偏推定量にはならない．標本 $X_1, \cdots, X_n$ があるとき，母分散 σ^2 の不偏推定量は，

$$U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2}$$

になる．ここで $\bar{X}$ は標本平均を表す．この U^2 を不偏分散という．不偏分散は平均，分散などの区間推定，検定の理論において，よく使われる概念である．

［2］一致推定量であること

標本数を大きくしていくと推定しようとする母数 θ に近づいていく統計量 Θ_n を，一致推定量（coincident estimator）という．すなわち，任意の $\varepsilon > 0$ に対して，

$$\lim_{n\to\infty} P(|\Theta_n - \theta| \geq \varepsilon) = 0.$$

標本平均 $\bar{X}$ は母平均 μ の一致推定量である．

$$\frac{X_1 + \cdots + X_n}{n-1}, \quad \frac{X_1 + \cdots + X_n}{n+3}$$

なども一致推定量である．しかしこれらは不偏推定量ではない．同様に考えて標本分散 S^2 も不偏分散 U^2 も母分散 σ^2 の一致推定量であることがわかる．

［3］有効推定量であること

2 つの不偏推定量を Θ_1，Θ_2 とする．

$$\mathrm{Var}(\Theta_1) < \mathrm{Var}(\Theta_2) \tag{3}$$

のとき，Θ_1 は Θ_2 より有効であるという．ここで Var は分散を表す．すべての不偏推定量の中で分散が最小のものがあれば，それを有効推定量（efficient estimator）という．

例えば，正規母集団 $N(\mu, \sigma^2)$ において，標本のメジアンは標本数が十分大きいときは，近似的に正規分布 $N\left(\mu, \frac{\sigma^2}{2n}\pi\right)$ に従うことが知られている．だから，標本のメジアンと標本平均はどちらも母平均の不偏推定量であり一致推定量である．しかし，標本のメジアンの分散は $\frac{\sigma^2\pi}{2n}$，標本平均の分散は $\frac{\sigma^2}{n}$ であるから，標本平均は標本のメジアンより有効であると言える．母集団が正規分布の場合は，標本平均，不偏分散は有効推定量である．

最尤法（さいゆうほう）

一般の母数について，上に挙げた［1］［2］［3］をなるべくよく満たすものを見つける方法としてよく使われるのが，最尤法である．これは統計ソフトの中でも非常によく用いられている方法である．

$X_1, \cdots, X_n$ を連続分布からの無作為標本とし，この分布の母数 θ に依存する確率密度関数を $f_\theta(x)$ とする．データ $x_1, \cdots, x_n$ を $X_1, \cdots, X_n$ の実現値とし，

$$L(x_1, \cdots, x_n; \theta) = f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n) \tag{4}$$

とおく．これは $X_1, \cdots, X_n$ の同時密度関数になる．$L(x_1, \cdots, x_n; \theta)$ を θ の関数と見たものを尤度関数という．尤度関数を最大にする $\theta = \hat{\theta}$ を θ の推定値とする．この考え方を最尤法（maximum likelihood estimation）という．実際の計算は両辺の対数をとって微分することにより最大値を求めるというやり方で行う．最尤法で得られた値 $\hat{\theta}$ を最尤推定値，それを実現値とする確率変数 $\Theta = \hat{\theta}(X_1, \cdots, X_n)$ を最尤推定量という．離散型の母集団についても同様である．

例 4　母集団が正規分布 $N(\mu, \sigma^2)$ であるとする．大きさ n の標本 $x_1, \cdots, x_n$ を使って，μ と σ^2 の最尤推定値を求めると，

$$\mu\text{の最尤推定値} = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{x},$$

$$\sigma^2\text{ の最尤推定値} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$$

となる．σ^2の最尤推定量は，標本分散に一致する．だからこれと先ほど述べたこととを合わせると，σ^2の最尤推定量は不偏推定量ではないことが分かる．このように，最尤推定量はすべての点推定の要請を満たしているわけではないが，よい推定量としての基準を少なくとも近似的に満たしていることが知られている．

最尤法は，母集団の分布は既知だがその母数（平均，分散など）が未知のとき，手元にある標本が得られるための最も“尤もらしい”母数を求める手法と言える．最小２乗法を使って求めた回帰係数は，データの誤差による散らばりが正規分布と仮定して最尤法で求めたものと一致する．このことは正規分布が自然な分布であることを改めて感じさせる．

4. 区間推定

標本から母平均μのとり得る範囲を推定することを考える．標本の平均を表す確率変数$\bar{X}$の分布はすべての推定，検定理論の基礎である．

命題５　母集団が正規分布$N(\mu, \sigma^2)$に従っていて，そこからn個の標本を取り出すことを考える．標本の各々の確率変数X_iは，互いに独立で母集団と同じ正規分布$N(\mu, \sigma^2)$に従うと考えられる．このとき標本平均を表す確率変数$\bar{X}$は$N\left(\mu, \frac{\sigma^2}{n}\right)$に従う．確率変数$\bar{X}$の分布とは，大きさ$n$の標本をいくつも作ったとき，それらの平均値の作る分布である．

推定，検定のほとんどの理論では，母集団分布を正規分布と仮定している．ごく大雑把に言えば，すべての分布はその数を増やせば正規分布に近づくと言えるからである（定理２ラプラスの定理，**図１**，定理３中心極限定理およびその説明参照）．それに推定，検定の理論においては，標本の平均をとることが基本的な考え方になるので，中心極限定理によりいっそう正規分布と見なせやすくなる．また正規分布は，標準化変換をすることにより標準正規分布$N(0, 1)$上ですべての議論ができるのでとても使いやすい．

これらの話において，母集団は無限母集団と見ている．いくらその病気の人が世界中にたくさんいてもその数は有限である．しかし有限として話をすると，母集団の個数が異なるとその都度別の式が必要となり不便である．だから確率統計の理論においては，数の多い有限を無限で近似することが非常に多い．

命題５において，変数を標準化することにより次の定理を得る．

定理６　正規母集団$\Omega : N(\mu, \sigma^2)$から大きさ$n$の標本を取り出し，それを$n$個の独立な確率変数$X_1, \cdots, X_n$と考える．その標本平均を$\bar{X}$とするとき，

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{5}$$

は標準正規分布$N(0, 1)$に従う．

今正規分布において，実数α（0.05 または 0.01 を使うことが多い）に対して，

$$P(-A_\alpha \leq Z \leq A_\alpha) = 1 - \alpha \tag{6}$$

なるA_αを考える．例えば，$A_{0.05} = 1.96$である．
(5) を (6) に代入して

$$P\left(-A_\alpha \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq A_\alpha\right) = 1 - \alpha \tag{7}$$

を得る．

標本から母平均μのとり得る範囲を推定したいので，(7) の括弧の中の$\bar{X}$を実現値$\bar{x}$で置き換えたものをμについて解くと，

$$\mathrm{P}\left(\bar{x} - A_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + A_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{8}$$

これより次の公式を得る．

公式７　母平均の推定（母分散既知のとき）　正規母集団Ωの分散はσ^2であるとする．Ωから抽出された大きさnの標本の標本平均が$\bar{x}$であるとき，母平均μの信頼度$1-\alpha$の信頼区間は

$$\bar{x} - A_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + A_\alpha \frac{\sigma}{\sqrt{n}}. \tag{9}$$

標本平均が同じでも散らばり（母分散）が大きければ信頼区間の幅は広がる．また情報量（標本数）が多ければその幅は絞られる．それは (9) に表れている．

“信頼度 95% の信頼区間”の意味は，正確に言うと次のようである．「標本をいくつもとったとすると，そこから計算できる信頼区間はその都度異なる．その中で，信頼区間が真の母平均を含んでいる確率が 95% である」．

文 献

1. 儀我真理子：確率・統計の基礎．2014; ムイスリ出版.
2. 薩摩順吉：確率・統計．1989; 岩波書店.
3. 澤田 昇，田澤新成：統計学の基礎と演習．2005; 共立出版.
4. 篠原昌彦：確率・統計．1989; 朝倉書店.
5. 丹後俊郎：新版医学への統計学．1993; 朝倉書店.
6. 東京大学教養学部統計学教室：自然科学の統計学．1993; 東京大学出版会.
7. 服部哲也：理工系の確率・統計．2005; 学術図書.
8. 宮原英夫，白鷹増男：医学統計学．1992; 朝倉書店.

（受付：2013 年 12 月 2 日）
（受理：2013 年 12 月 26 日）