推測統計 [数理統計学]I

石綿 元

第 6 回講義

目次

1　推測統計概論

1.1　統計学の考え方

1.1.1　統計的思考

■演繹的思考　数学による証明や計算の解法などは、「いつ」「どこで」「だれが」求めても同じである。そこでの思考過程はすべて連続的であり、思考の不連続な部分は一点も存在しない。代表的な例として３段論法があげられるが、思考が別のステージへ遷移する場合においては、それまでの思考のすべての積み上げと組み合わせのもとで行われる。それが数学が厳密であることのよりどころとするものである。このような思考法を演繹的思考と呼ぶ。完全完璧な論理である。

■帰納的思考　演繹的思考の対義語として存在するのが帰納的思考である。すなわち、思考の連続性が必ずしも保証されていない。科学的思考は、本来的に帰納的思考を行う。それは、個別の事象の発見を、利用できるような一般化した形の知識として残していく活動が本来的な科学の活動だからである。代表的な例として万有引力の発見で考えてみよう。アイザック・ニュートンは、「リンゴが木から落ちる。(個別の事象)」から「万有引力の法則（一般化された知識)」を「はっと思いついた（帰納的思考による推測)」という。もっと言うと、生命が生存のために本質的に備えているであろう思考法である。例えば、古代人が海辺に引越してきて初めて魚を穫り、風船のように膨らむ魚を獲得したとしよう。持ってかえった人の一家が翌朝全滅していた（個別の事象）ならば、それを見た人は、その魚を食べようとは思わない（一般化された知識）であろう。しかし、この例の場合に原因は、おそらくその魚によるものであろう（帰納的思考による推測）が、もしかしたら一酸化炭素中毒による事故であったかもしれない。つまり、不確実性を多分に含み、原因と結果が必ずしも１対１の関係を保持しているとは限らない。

　このような思考法を帰納的思考と呼び、不確実な現実ではいつも求められる思考であり、それが実質的な科学的思考でもある。

■統計的思考　統計的思考では、帰納的思考と演繹的思考をスムーズにつなげ、不確実な現実において、帰納的思考の結論である一般化された知識を求める際に、データにより「数」を介在させ演繹的思考を利用して、帰納的思考による結論をより補強または修正するような結論を求める。このように扱うことにより、より多くの社会的同意（ここでいう社会的同意とは、科学の分野では論文など成果を発表するということ）を得られるように整理された結論が得られる。統計的思考を用いたとしても、不確実性のある問題に完全完璧な結果を求めることはできないが、よりベターな解を求めることができるため、その後の戦略としてよりベストな判断に結びつけられるのである。

1.1.2　記述統計と推測統計

　統計学の考え方は、記述統計と推測統計に大別することができる。

■記述統計　いま、対象とする集団について、調べたい事柄についてデータがすべて得られているならば、「第１回講義　**データを代表値で読む**」で解説したように、正確な平均や分散を得ることができ、「第１回講義　**データの視覚化**」で扱ったような視覚化を行い、対象とする集団がどのようなものかを調べていく。このような方法の体系を記述統計と呼ぶ。５年に一度行われる「国勢調査」は、日本国民全員についてすべてデータを得ることを目的に行われる。このような調べたい対象のデータをすべて得ることを悉皆調査または全数調査と

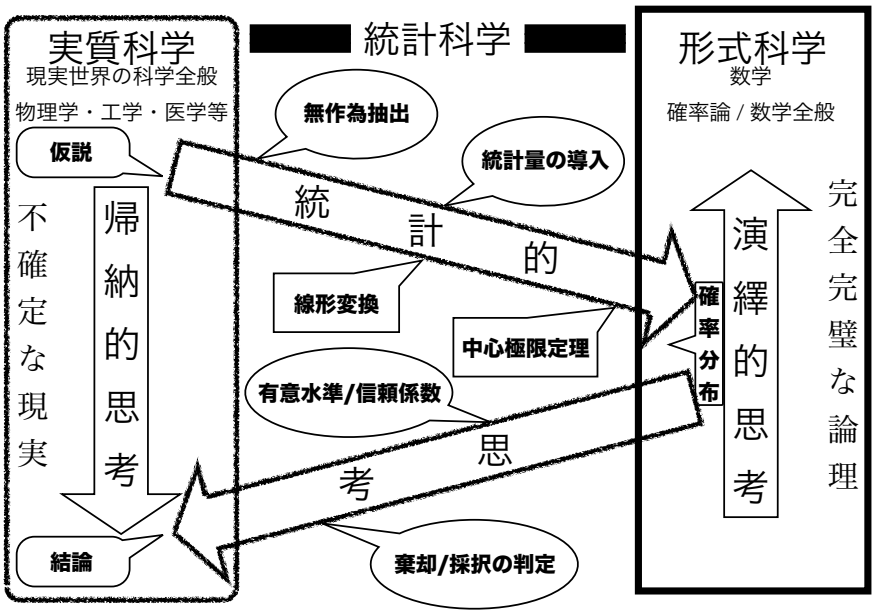

図 1　統計的思考の概要

よぶ。

■推測統計　一方、対象とする集団について、調べたい事柄についてのデータがすべて得られない場合には、「確率論」などの数学を利用して、得られうるデータをもとにして対象集団全体について推測を行う。このような方法の体系を推測統計と呼ぶ。本章以降の講義では推測統計を扱う。

1.1.3　尺度水準

ところで、第 1 回講義の記述統計において説明した通り、統計では質的データも量的データも扱うのであった。これらデータを得る際に、量的データでも質的データでも数値と対応させる必要がある。その決まりのことを尺度と呼ぶ。つまり「ものさし」である。

この尺度を大きく分類すると一般に 4 つの尺度水準に分類され、それぞれのデータの性質において可能となる演算が異なる。

1. 質的データ
   - 名義尺度：同じ値であるかないかだけが問題となるデータであって、たいていの場合 0 か 1 もしくは 1 か 2 を対応させることができる。アンケートにおける男女別の記入のような場合のデータが該当する。この場合、四則演算、平均・分散といった値に意味がない。
   - 順序尺度：同じ値であるかないかに加えて、大小関係があり順番について意味をもつデータであって、順序を数値と当てはめることができる。アンケートにおける嗜好順位の調査などのデータがあげられる。
2. 量的データ
   - 間隔尺度：同じ値であるかないか、大小関係、およびデータ同士の和・差について意味をもつデー

タであって、間隔をそろえて分類できる場合のデータである。たとえば、５段階評価の成績データなどがあげられる。

- 比率尺度：原点が定まっており、同じ値であるかないか、大小関係、四則演算のすべてが可能で、意味を持つデータ。ほとんどの物理量が該当する。

1.1.4　統計学は「尺度の科学」

そもそも尺度を作り上げ、社会的同意を得るための規準を提供すること、それこそが本質的な統計学の持つ構造である。「統計」とは、その名の通り「統治するための計り」である。それは、英語であっても同様で「Statistics」とは、「The United States of America」での「States」と由来を同じくし、やはり政府・国家・州といった統治するという意味を持っている。

推測統計の大部分では、確率分布の面積・対応する横軸の値などは、数表で与えられる。例えていうならば、時を計る道具としての時計において、数表は時計の文字盤のような役割を果たす。現実に利用していく場合には、統計ソフト「R」などを利用し、コンピュータで計算していくことになる。いずれにしても、目的の数値はそれほど大変な計算をすることなく得られるが、その数値の持つ意味や、その数値をどのように結論に結びつけるかの解釈が統計では重要になってくる。

1.2　ランダムサンプリング

1.2.1　標本抽出

■母集団と無作為抽出　いま知りたい対象についての数値データの全体集合を母集団という。母集団の要素が少数ならば全数調査が簡単にできる。この場合、記述統計で対応できる。

母集団の要素が大きい、特に無限に大きいと考える場合、どうすれば対象とするもの全体の性質を知ることができるか？を考えていく。

母集団からランダムに取り出した要素を標本という。

ランダムに取り出すとは、母集団の部分集合のどの組もその要素が標本に選ばれる確率が同じ場合をいい、このとき、無作為抽出という。母集団から標本を無作為抽出することを標本抽出という。

標本抽出する ＝ 無作為抽出であること

無作為抽出を行う場合、乱数表を用いるとよい。乱数表とは、どの数値も全く等確率で出現するように数値をランダムに並べたものであり、フィッシャーらが発表した６表が有名である。また、コンピューター上において純粋な乱数を発生させることは難しく、現代では放射性崩壊や半導体の熱雑音などハードウェアにおいて物理的に発生させる乱数ボードなども開発されている。

1.2.2　推測統計の方法

推測統計の考え方を俯瞰しておく。

おおざっぱに言うと、無作為抽出した標本を用いて母集団を推定するために、各種確率分布を知りたい母数に応じて使い分け、その確率分布の性質に応じて標本から母数を推定し、母集団について知ろうというものである。

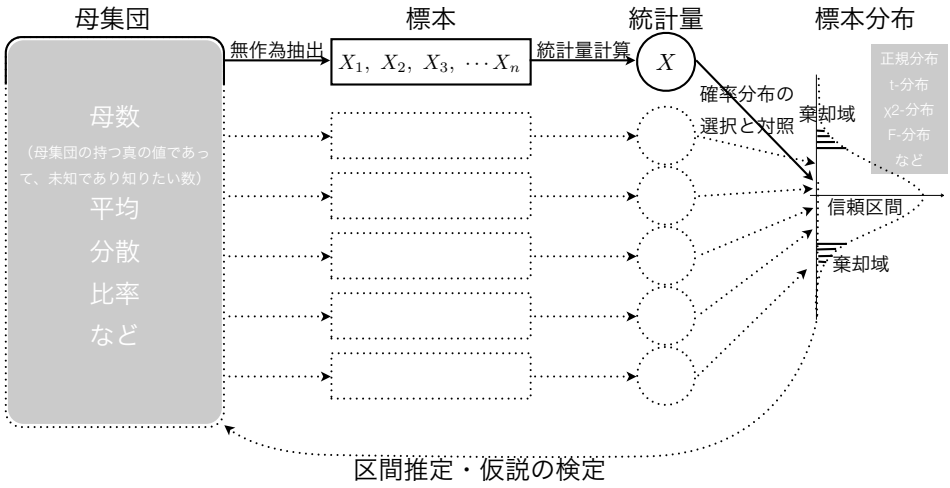

図 2　一般的な推測統計を理解するための概念図

■標本抽出と中心極限定理　母集団を確率変数 X を持つ確率分布として扱うとする。このとき、この母集団から無作為抽出した標本は、当然、母集団と同じ確率分布に従っており、かつ、無作為抽出した各標本はそれぞれ独立である。したがって、それぞれの標本について確率変数 X_i とすると、母集団の確率変数 X と同一の確率分布に独立に従う n 個の標本による各確率変数 $X_1,\ X_2,\ X_3,\ \cdots X_n$ である。すなわち、「n 個のまったく独立な確率変数 $X_1,\ X_2,\ X_3,\ \cdots X_n$ が、すべて独立に確率変数 X と同一の確率分布に従う」ので、n が十分大きいとき、大数の法則、さらには中心極限定理を適用することができる。また、無作為抽出した標本を無作為標本といい、無作為標本の確率変数がつくる確率分布を標本分布という。

1.3　母集団と標本分布

1.3.1　母数

母集団の確率分布を特徴付ける量を母数といい、一般には未知である。例としては、母集団 X の平均 μ、分散 σ^2 など。

1.3.2　統計量

統計量とは、標本だけで定義される量を統計量という。また、統計量には未知のパラメータを含まない。例としては、標本平均、標本の分散、標本の標準偏差、標本のメディアン、標本の最大値、標本の最小値など。

■標本平均　無作為抽出した標本 $X_1,\ X_2,\ X_3,\ \cdots X_n$ についての平均を標本平均といい、次で定義する。

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

n が十分大きいならば、確率変数 X が従う分布によらず、総和の平均 $\bar{X}$ は平均 μ、分散 $\frac{\sigma^2}{n}$ の

正規分布 $N\left(\mu,\ \frac{\sigma^2}{n}\right)$ に近づく。

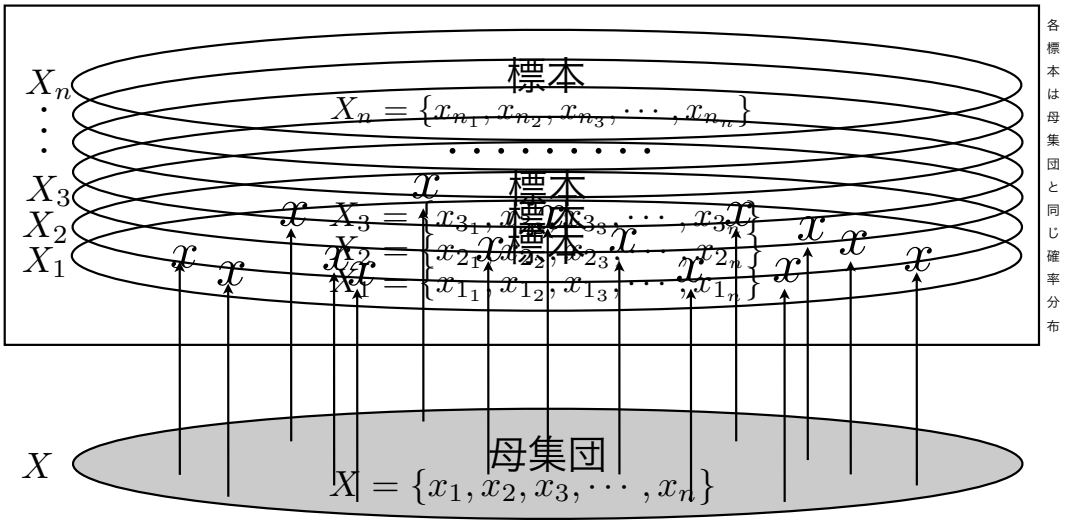

図 3　標本抽出と中心極限定理

■不偏標本分散　無作為抽出した標本 $X_1,\ X_2,\ X_3,\ \cdots X_n$ についての分散を不偏標本分散といい、次で定義する。

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

これ以後の推測統計では、通常、標本分散というときには、この不偏標本分散を用いていく。

■標本の分散について　記述統計で導入した分散と同様の定義を用いた標本の分散も統計量であるが、一般に推測統計の分野では、不偏性を備えていないと考える。不偏性については、この後の節で解説する。

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

ただし、「第１３回講義　**情報量規準**」の講義回で解説する最尤推定量を用いるならば、n で割る分散の定義（不偏でない標本分散）と同じ値が推定量として導かれ、この方がもっともらしいととらえる（この定義を尺度として利用することにも一定程度の根拠がある）。

1.4　パラメトリックモデルと推定量

1.4.1　正規分布の発見と利用

18 世紀以降多くの科学者により正規分布の導出はなされているが、中でも重要なのは、ドイツの物理学者ヨハン・カール・フリードリヒ・ガウスによる天体観測の誤差評価により、正規分布が導きだされたことである。このガウスの誤差分布について確認しておく。

まず、天体観測による真の値とのずれについて考察すると、真の値を a_0 として、観測回ごとの観測値を a_i とするなら、各観測回における誤差を $a_i - a_0 = x_i$ と表すことにする。この x_i を観測値とする X を確率変数とする確率分布を求める。

いま、確率変数 X の作る確率分布を $f(x)$ とするとき、

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{1}$$

である。確率分布の任意区間の積分が、その区間の確率変数が取りうる確率を表すから、このとき変数 x_i との微少変化量 $x_i + dx$ とすると、その間 $[x_i,\ dx + x_i]$ の積分は、その区間の確率を表す。ところで、この区間 $[x_i,\ dx + x_i]$ は、微少幅 $dx = x_i + dx - x_i$ で、高さ $f(x_i)$ だから、$f(x)$ についての積分はリーマン和の短冊一つと考えれば、

$$P(dx) = f(x_i)dx$$

いま、n 回観測したとして、その観測値は $a_1, a_2, a_3, \cdots, a_n$ であるが、すべて独立に得られたものである。したがって、$x_1 = a_1 - a_0,\ x_2 = a_2 - a_0,\ \cdots x_n = a_n - a_0$ は、すべて独立であるから、各 x_i ごとに $P(dx)$ を考えると、

$$\begin{aligned}
\{P(dx)\}^n &= \{f(x_1)dx\} \cdot \{f(x_2)dx\} \cdots \{f(x_n)dx\} \\
&= \{f(a_1 - a_0)dx\} \cdot \{f(a_2 - a_0)dx\} \cdots \{f(a_n - a_0)dx\} \\
&= \{f(a_1 - a_0)\} \cdot \{f(a_2 - a_0)\} \cdots \{f(a_n - a_0)\} (dx)^n
\end{aligned} \tag{2}$$

ここで、$a_i - a_0$ は、プラスも取ればマイナスも取る微少量で 0 近辺を行き来する。つまり、$P(dx)$ とは、$f(0)$ を取るとき、最大と考えてよく、それは、真の値 a_0 を取る確率を表していることに他ならない。よって、観測回数 n を十分大きくするとき、a_0 で $P(dx)$ は最大値を取るので、

$$\frac{dP(dx)}{da_0} = 0 \tag{3}$$

が成り立っていなければならない。ここで、式 (2) の両辺に対数を取ると、

$$\begin{aligned}
\log\left[\{P(dx)\}^n\right] &= \log\left[\{f(a_1 - a_0)\} \cdot \{f(a_2 - a_0)\} \cdots \{f(a_n - a_0)\} (dx)^n\right] \\
n \log\{P(dx)\} &= \log\left[\{f(a_1 - a_0)\} \cdot \{f(a_2 - a_0)\} \cdots \{f(a_n - a_0)\}\right] + n\log(dx) \\
&= \log\{f(a_1 - a_0)\} + \log\{f(a_2 - a_0)\} + \cdots + \log\{f(a_n - a_0)\} + n\log(dx)
\end{aligned} \tag{4}$$

この式 (4) を両辺 a_0 で微分すると、

$$\begin{aligned}
\frac{n}{P(dx)} \frac{dP(dx)}{da_0} &= -\frac{f'(a_1 - a_0)}{f(a_1 - a_0)} - \frac{f'(a_2 - a_0)}{f(a_2 - a_0)} \cdots - \frac{f'(a_n - a_0)}{f(a_n - a_0)} + 0 \\
&= -\sum_{i=1}^{n} \frac{f'(a_i - a_0)}{f(a_i - a_0)}
\end{aligned}$$

つまり、

$$-\frac{n}{P(dx)} \frac{dP(dx)}{da_0} = \sum_{i=1}^{n} \frac{f'(a_i - a_0)}{f(a_i - a_0)}$$

ここで、式 (3) が成立している訳だから、

$$-\frac{n}{P(dx)}\frac{dP(dx)}{da_0}=0$$

よって、

$$\sum_{i=1}^{n}\frac{f'(a_i-a_0)}{f(a_i-a_0)}=0 \tag{5}$$

である。ここで、

$$\frac{f'(a_i-a_0)}{f(a_i-a_0)}\equiv g(x_i) \tag{6}$$

とおく。よって、式 (5) の条件は、

$$\sum_{i=1}^{n}\frac{f'(a_i-a_0)}{f(a_i-a_0)}=\sum_{i=1}^{n}g(x_i)=0 \tag{7}$$

と表せる。

ところで、x_i とは、誤差であり、a_i-a_0 であるから、当然、プラスも取ればマイナスも取る微少量で 0 近辺を行き来する値であって、

$$\sum_{i=1}^{n}x_i=0 \tag{8}$$

が成立している。すなわち、

$$x_1+x_2+\cdots+x_n=0$$

であるから、

$$x_n=-(x_1+x_2+\cdots+x_{n-1}) \tag{9}$$

とも表る。

一方、式 (7) についても具体的に書き表すと、

$$\sum_{i=1}^{n}g(x_i)=0$$
$$g(x_1)+g(x_2)+\cdots+g(x_n)=0$$

であるから、式 (9) を適用すると、

$$g(x_1)+g(x_2)+\cdots+g(-(x_1+x_2+\cdots+x_{n-1}))=0$$

これを x_1 で微分すると、

$$g'(x_1)+0+\cdots+0+g'(x_n)\cdot\left(\frac{dx_n}{dx_1}\right)=0$$

いま、左辺最終項は、合成関数の微分であり、$x_n=-(x_1+x_2+\cdots+x_{n-1})$ より $\frac{dx_n}{dx_1}=-1$ なので、

$$g'(x_1)=g'(x_n)$$

同様に

$$g(x_1) + g(x_2) + \cdots + g(-(x_1 + x_2 + \cdots + x_{n-1})) = 0$$

これを x_2 で微分すると、

$$0 + g'(x_2) + 0 + \cdots + 0 + g'(x_n) \cdot \left(\frac{dx_n}{dx_2} \right) = 0$$

いま、左辺最終項は、合成関数の微分であり、$x_n = -(x_1 + x_2 + \cdots + x_{n-1})$ より $\frac{dx_n}{dx_2} = -1$ なので、

$$g'(x_2) = g'(x_n)$$

すなわち、

$$g'(x_1) = g'(x_2) = \cdots = g'(x_n)$$

であるから、$g'(x_i)$ は、変数 x_i によらない定数である。したがって、定数 α, β を用いて、

$$g(x) = \alpha x + \beta \tag{10}$$

と表せる。いま、式 (7)、式 (8) より $\sum_{i=1}^{n} g(x_i) = 0$ および $\sum_{i=1}^{n} x_i = 0$ であったから、式 (10) を用いて、

$$\begin{aligned} \sum_{i=1}^{n} g(x_i) &= \alpha x_1 + \beta + \alpha x_2 + \beta + \cdots + \alpha x_n + \beta = 0 \\ &= \alpha \left(x_1 + x_2 + \cdots + x_n \right) + n\beta = 0 \\ &= \alpha \left(\sum_{i=1}^{n} x_i \right) + n\beta = 0 \\ &= 0 + n\beta = 0 \end{aligned}$$

いま、観測回数 n は当然ながら $n \neq 0$ だから、

$$\beta = 0$$

よって、

$$g(x) = \alpha x$$

ここで、$g(x)$ を定義した式 (6) から元に戻すと、

$$g(x_i) = \frac{f'(a_i - a_0)}{f(a_i - a_0)} = \alpha x_i$$

また、$a_i - a_0 = x_i$ であったので、

$$\frac{f'(x)}{f(x)} = \alpha x$$

である。よって、

$$f'(x) = \alpha x f(x)$$

ここで、変数分離形として、

$$\begin{aligned}\frac{dy}{dx} &= \alpha xy \\ \int \frac{1}{y}dy &= \int \alpha x dx + \gamma \\ \log y &= \alpha \frac{1}{2}x^2 + \gamma \\ y &= e^{\frac{\alpha x^2}{2}+\gamma} \\ &= e^{\gamma} e^{\frac{\alpha x^2}{2}}\end{aligned}$$

ここで、$C \equiv e^{\gamma}$ とおくと、

$$f(x) = Ce^{\frac{\alpha x^2}{2}} \tag{11}$$

が求まる。この $f(x)$ が求めたい確率分布であるが、いま、誤差が従う分布を考えていることを忘れてはならない。誤差ということは、重大なインシデントは滅多に起こらない。すなわち、大きな誤差が起こることは非常に稀で、そんな確率は小さいはずである。したがって、$\alpha < 0$ であり、かつ、十分小さい値とするために $\alpha = -\frac{1}{\sigma^2}$ と仮定しておく。そうすることによって、式 (11) に代入すると、

$$f(x) = Ce^{-\frac{x^2}{2\sigma^2}} \tag{12}$$

である。さらに、式 (1) で確認した通り、この $f(x)$ は確率密度関数であるから、

$$\int_{-\infty}^{\infty} Ce^{-\frac{x^2}{2\sigma^2}} dx = 1 \tag{13}$$

これによって、未知数 C を確定していく。ここで、$\frac{x}{\sqrt{2}\sigma} = t$ とおいて変数変換すると、$\sqrt{2}\sigma dt = dx$ で、積分区間は変わらないから、

$$\begin{aligned}\int_{-\infty}^{\infty} Ce^{-\frac{x^2}{2\sigma^2}} dx &= \int_{-\infty}^{\infty} Ce^{-t^2} \cdot \sqrt{2}\sigma dt = 1 \\ &= \sqrt{2}\sigma \cdot C \int_{-\infty}^{\infty} e^{-t^2} dt = 1 \\ &\text{ここで、確率積分} \int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}\text{を用いると} \\ &= \sqrt{2}\sigma \cdot C \cdot \sqrt{\pi} = 1\end{aligned}$$

よって、

$$C = \frac{1}{\sqrt{2}\sqrt{\pi}\sigma}$$

である。

よって、求める誤差分布の確率密度関数は、

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}$$

である。これは、平均 0、分散 σ^2 の正規分布 $\mathrm{N}(0, \sigma^2)$ に他ならない。

実験観測の誤差や、製造工程における不良品の製造などは、天体観測の測定誤差と同様そもそも意図しない小さい値である。このような誤差が持つ分布は、多くの場合、はじめから正規分布に従うと考えてよい。

また、２項分布の正規近似など中心極限定理を介した多数の標本の**平均**が従う分布は正規分布である。

さらに、人体を含む生命体の個体差や能力差などについてのデータは、そもそも対象集団が正規分布に従っていると、あらかじめ仮定した上で考えていく。

ただし、世の中のすべてのデータはいつも正規分布に従うと勘違いしてはいけない。現実の世界はそんな単純なものではないことにも留意しておかなければならない。

1.4.2　パラメトリックモデル

いま、母集団を知ることが目的であり、母数は未知である。そこで、母集団を確率分布として扱ったとき、その確率分布を特徴付ける量である母数を知りたい。もしくは、母数を推定することで確率分布を知りたい。そこで、標本を用いる。

つまり、無作為抽出した標本 $X_1,\ X_2,\ X_3,\ \cdots X_n$ によって、母集団の確率変数 X の確率分布を推定する。このとき、X の確率分布は事前に仮定しておく。一般には正規分布に従っているものとして扱う。

また、中心極限定理により、標本数が多い場合には、その標本**平均**がつくる分布は正規分布に従うのであった。

このように**事前に確率分布を仮定できる場合をパラメトリックモデルという**。以降、「第１０回講義　**分散分析**」の講義回まではすべてパラメトリックモデルを扱う。また、事前に確率分布を仮定しない場合の一般的な推測統計は、「第１１回講義　**母数によらない方法**」で扱う。

1.4.3　推定量

確率分布が仮定されてもその確率分布に特徴的な量がわからないと分布がわかったことにはならない。つまり、確率密度関数を具体的に表記できない。したがって、仮定された母集団の確率分布に特徴的な量を標本から推定する。このとき、特徴的な量を推定するために標本から求めた統計量を推定量という。たとえば、母集団が正規分布 $\mathrm{N}(\mu, \sigma^2)$ に従っていると仮定したとき、特徴的な量とは平均 μ と分散 σ^2 である。

また、記号として**推定量 $\hat{\theta}$ に対して母数 θ と表記する**ことを導入しておく。

1.5　推定量の一致性と不偏性

1.5.1　一致性

推定量 $\hat{\theta}$ について大数の法則をはじめとする極限定理を適用すると母数 θ に近づくとするとき、一致推定量という。たとえば、標本平均 $\bar{X}$ は標本数が十分多ければ大数の法則によって母平均 μ に近づく。

1.5.2　不偏性

推定量の期待値が母数と一致しているものを不偏推定量という。

$$E[\hat{\theta}] = \theta$$

■標本平均の期待値　標本平均 $\bar{X}$ の平均について考えると、

$$E[\hat{\mu}] = E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}n\mu = \mu$$

ここで、μ は母集団の平均である。よって、標本平均は不偏性を備えている。

■標本平均の分散　標本平均 $\bar{X}$ の分散について考えると、

$$V[\bar{X}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{i=1}^{n} X_i - E\left[\sum_{i=1}^{n} X_i\right]\right)^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])^2\right) = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

ここで、σ^2 は母集団の分散である。

■不偏標本分散の期待値　不偏標本分散 s^2 の平均について考える。ここで、偏差平方和を $T_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ とおく。

$$\begin{aligned}
T_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}\left((x_i - \mu) - (\bar{x} - \mu)\right)^2 \\
&= \sum_{i=1}^{n}\left((x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2\right) \\
&= \sum_{i=1}^{n}(x_i - \mu)^2 - 2\sum_{i=1}^{n}(x_i - \mu)(\bar{x} - \mu) + n(\bar{x} - \mu)^2 \\
&= \sum_{i=1}^{n}(x_i - \mu)^2 - 2n(\bar{x} - \mu)^2 + n(\bar{x} - \mu)^2 \\
&= \sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2
\end{aligned}$$

としておく。これを用いて、

$$\begin{aligned}
E[\hat{\sigma^2}] = E[s^2] &= E\left[\frac{1}{n-1}T_{XX}\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}E\left[(X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}E\left[(X_i - \mu)^2\right] - nE\left[(\bar{X} - \mu)^2\right]\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}V[X_i] - nV[\bar{X}]\right) \quad \text{標本平均の分散なので、} \\
&= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) \\
&= \frac{1}{n-1}(n-1)\sigma^2 \\
&= \sigma^2
\end{aligned}$$

ここで、σ^2 は母集団の分散である。よって不偏標本分散である $\frac{1}{n-1}\sum(X_i - \bar{X})^2$ は不偏推定量であり、$\frac{1}{n}\sum(X_i - \bar{X})^2$ は不偏推定量とはならない。なお、この不偏性確認のための式変形において、途中標本平均の分散を利用している。$n-1$ で割ることについて、n で割るよりも 1 減らしておくことは、不偏標本分散の定義には推定量の標本平均を用いていることから、自由度 n が 1 固定されていると考えることと同意である。

索引