現代数理統計学の基礎
6 章 統計的推定

中村俊

2020 年 7 月 8 日

0　概要

データを分析する際、一般的に分析対象であるデータの中身を把握するための手段として基礎的な集計 (以下、基礎集計) を実施することが多いが、単純にいくつかの特徴量を縦軸ないしは横軸に置き、サンプル間における分布を確認するだけでは、定性的には「ばらついている」「分布が偏っている」などが言えるが、分布を陽に可視化しなくても、分析対象であるデータを定量的に解釈する指標が存在する。大量のサンプルデータをまとめて論述することを「データの縮約」といい、縮約の表現、つまり、多くのサンプルからなるデータを要約かつ定量的に表現するための指標が「十分統計量」である。

上述した十分統計量にはいくつか種類があり、本章では我々が普段手にするサンプルデータからそれらを推定する手法を 3 つ紹介する。詳細は後述するが、十分統計量の推定方法として、モーメント法、最尤法、ベイズ法が存在する。本稿では、これら 3 つの十分統計量の推定方法だけでなく、これらの中から最適な手法を選ぶための評価基準である、分散の最小化やクラメール・ラオの不等式についても述べる。

1　統計的推測

基本的に縮約表現として、統計的に推定を行う対象となるものは、データの母集団全体が持っているパラメーターである。しかしこのパラメーターを推定するために必要になる情報がいくつかあり、その 1 つとして挙げられるのが、「十分統計量」である。

ここでは、「十分統計量」からパラメーターを推定するという「統計的推定」の考え方や、そもそもの縮約表現としての「十分統計量」や「因子分解定理」について説明する。

1.1　統計的推測の考え方

「何を推測するか」という観点で述べると、手元にあるデータから「母集団の確率分布」を推測することが統計的推測の考え方である。なので、手元にあるデータは全データではなく、あくまで「母集団から標本抽出された一部」ということが、統計的推測について言及する上での前提条件となる。

ここで統計的推定を行う際に問題となるのが、標本抽出された異なる 2 群のデータから推定された何かしらの母集団の分布に関するパラメーターの値の差は*統計的に有意な差と言えるのだろうか？*ということである。

この問題を考えるため例として、手元にあるデータで日本全国における内閣支持率を算出したい場合を考える (繰り返しになるが、日本の全都道府県のデータは何かしらの理由で手に入らないとする)。日本全国のデータのうち、標本抽出されたデータ A 群と B 群を考える。

先月は、A 群のデータから内閣支持率[*1]を算出し、今月は B 群のデータから内閣支持率を算出する場合を考える。この場合、それらの内閣支持率の増減は、A 群と B 群と異なるデータを用いて算出したのが原因であり、本質的な内閣支持率の増減を意味しているとは限らず、算出元のデータが異なることによる「誤差」という可能性も考えられる。

この「誤差」=「データの揺れ」に対応するために、内閣支持率を算出するための日本全体のデータが「ある確率分布」にしたがっていると仮定する。「データが確率分布に従っている」とは、その確率分布の有するパラメーターを θ とすると、以下のような関数から生成されたデータであると言い換えることができる。、母集団からサンプリングされた実際のデータの値である。

$$f(x \mid \theta) \tag{1}$$

そして、手元にある A 群（もしくは B 群）のデータから、上記で定式化される日本全体のデータの分布がもつパラメーター「統計的点推定」もしくは「推定」と呼ぶ。

上記の確率分布の関数系が既知の場合にパラメーターのみを推定するモデルをパラメトリックモデルと呼び、パラメーターだけでなく関数系も未知な場合の推定モデルをノンパラメトリックモデルと呼ぶ。

ここでいう関数系とは、データの分布として、正規分布なのか、ベーター分布なのか、ポアソン分布なのか、などのデータの分布形状の仮定関数のことである。

関数系は不明だが、平均値や分散などのパラメーターが部分的に仮定できる場合は、セミパラメトリックモデルと呼ぶ。

1.2 データの縮約

データが大量にある場合、そのデータ全体を見渡す指標として、「十分統計量」なるものを算出することで、そのデータを縮約できることは本稿の概要でも述べた通りである。ここではまず、データの縮約のための「十分統計量」の定義について述べていく。

縮約とは、そのデータの「概要」とも言えるが、なるべく元の情報[*2]を多く含んでいる方が望ましく、かつシンプルであることが望ましい。当然、縮約した結果、元の情報とは全然異なる情報に変わってしまうことは、そもそも縮約になっていない。具体的に言えば、「統計量を用いてデータの概要を述べた結果は、その統計量を使わなくともデータから言えることと同じ」でないと縮約にはなっていない。「十分統計量」とはそのような統計量のことである。

例えば、統計量として「サンプルの合計」を考える。上述した「統計量を用いてデータの概要を述べた結果は、その統計量を使わなくともデータから言えることと同じ」というのは、「母集団に関する情報を失っていない」とも言える。母集団に関する情報=母集団のパラメーター θ として、その統計量 t が与えられた時のサンプルの生成確率（条件付き確率）

$$P(X = x \mid T(X) = t) \tag{2}$$

[*1] あくまで一例のため、内閣支持率の計算方法についてはここでは論じない

[*2] 正しくはフィッシャー情報量。詳細は 6.3 章で述べる。

が θ に依存しないような統計量 t が「元のデータの情報を失っていない、使うに値する十分な統計量」＝「十分統計量」と言える。

1.3 因子分解定理

十分統計量について、概念的な説明は以上で述べた通りであるが、ここでは数学的な導出へと帰着させるための重要な定理である「因子分解定理[*3]」について述べる。

以下テキストの抜粋である。

因子分解定理

$T(\boldsymbol{X})$ が θ の十分統計量であるための必要十分条件は、$\boldsymbol{X} = (X_1, \cdots, X_n)$ の同時確率関数もしくは同時確率密度関数 $f(x_1, \cdots, x_n \mid \theta)$ が θ に依存する部分とそうでない部分に分解でき、θ に依存する部分は $T(\cdot)$ を通してのみ $\boldsymbol{x}$ に依存する。すなわち

$$f(x_1, \cdots, x_n \mid \theta) = h(\boldsymbol{x})g(T(\boldsymbol{x}) \mid \theta) \tag{3}$$

と表されることである。

この定理の重要な点は、必要十分条件についての定理だということである。つまり式 (3) のように、確率密度関数が式変形さえできれば、$T(\boldsymbol{X})$ が十分であるということだけでなく、$T(\boldsymbol{X}) = t$ を満たす、$\boldsymbol{X} = (X_1, \cdots, X_n)$ もまた十分統計量であることを意味しているということが重要。

因子分解定理の使い方として、例題を解く。

例題 1: $X_1, \cdots, X_n$ が独立にポアソン分布 $P_o(\theta)$ に従う時、θ の十分統計量を求めよ。

解答

同時確率を考えるので相乗が式中に出現している。これを相乗が登場しない形式に式変形する。

$$\begin{aligned} f(x_1, \cdots, x_n \mid \theta) &= \prod_{i=1}^{n} \frac{e^{\theta}\theta^{x_i}}{x_i!} \\ &= \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}}{x_1!x_2!\cdots x_n!} \end{aligned}$$

となり、分母の $\frac{1}{x_1!x_2!\cdots x_n!}$ は θ を含まないため、因子分解定理を用いると、$\sum_{i=1}^{n} x_i$ は十分統計量である。$\sum_{i=1}^{n} x_i$ が十分統計量ということは、$\sum_{i=1}^{n} x_i$ が一意に定まれば求まるその他の統計量もまた十分統計量である。つまり、$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}$ も十分統計量である。

因子分解の利便性を体感するためにもう 1 つ、例題を確認する。

例題 2: https://www.toukei-kentei.jp/wp-content/uploads/201911grade1suri.pdf (2019 年の統計検定 1 級 統計数理 問題 3-1) テキストの例題 6.4 と問われていることは同じ。

[*3] 文献によっては「Neyman の因子分解定理」と記載されていることもある

解答

以下テキストの抜粋である。

以上の例のように、因子分解定理を用いることで、「十分統計量」を求めることと、「十分統計量」であることを示すことはどちらも容易に済ませることができる。これは上述の通り、因子分解定理が標本 $\boldsymbol{X}$ の統計量 $T(\boldsymbol{X})$ に関して「必要十分条件」を満たしているからである。

1.4 因子分解定理の指数型分布族への適用

確率密度関数が以下の形式で表現できる確率分布（正規分布、ガンマ分布、2 項分布[*4]、ポアソン分布など）は、指数型分布族呼ばれる。

$$f(x \mid \theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\}$$

ここで、$w_i(\cdot)$ は実数値関数、$t_i(x)$ はデータ x が与えられた時の統計量である（結果的には十分統計量になる）。

上式で表せられる指数型分布族に対しては、これまでの議論をより一般的に拡張できることが知られている。例題同様、サンプリングデータ $\boldsymbol{X}$ における同時確率分布を考えると、

$$f(\boldsymbol{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} h(x)c(\boldsymbol{\theta})\exp\left\{\sum_{j=1}^{k} w_j(\boldsymbol{\theta})\sum_{i=1}^{n} t_j(x_i)\right\}$$

のように表せられ、因子分解定理により、$\sum_{i=1}^{n} t_j(x_i)$ が $\boldsymbol{\theta}$ に対する十分統計量になる。普段何気なく母集団における平均値 μ や σ^2 の統計量として、標本抽出されたデータ（我々が普段扱う手元のデータ）の平均値 $(\bar{X})$ や分散 (S^2) を使用しているのは、上述した定理によりその「十分性」が「必要十分条件」として満たされているからである。ここで論じたことは、文献 [**?**] では「指数型分布族の完備性」として詳細に述べられている。

2 点推定量の導出方法

これまでの議論は、点推定された統計量の縮約表現としての正しさについて述べてきたが、ここではその統計量の推定方法である、モーメント法、最尤法、ベイス法（ベイズ推定）の 3 つの手法について述べていく。

2.1 モーメント法

推定される統計量は、どんなものが考えられるだろうか？代表的な統計値（$\neq$ 推定される統計量）としては、平均値、分散、歪度、尖度などが考えられるが、それらの各定義の式より、モーメント $E[X], E[X^2], E[X^3] \cdots, E[X^k]$ （kは任意の自然数）値が簡単な代数方程式から算出できることには意義がある。

[*4] 詳細は Appendix を参照

X $f(x \mid \boldsymbol{\theta})$ なる確率変数を考えた場合、未知パラメーター θ に関する推定量は、$E[X] = \mu_1'(\boldsymbol{\theta}), E[X^2] = \mu_2'(\boldsymbol{\theta}), E[X^3] = \mu_3'(\boldsymbol{\theta}) \cdots, E[X^k] = \mu_k'(\boldsymbol{\theta})$ とおける。

詳細は本章の範囲を超えるため、割愛するが、$\frac{1}{n}\sum_{i=1}^{n} X_i^k$ は、大数の法則により、$E[X^k]$ に確率収束することが知られているので、

未知パラメーターの推定量に関する各方程式の左辺を置き換えると、

$$
\begin{aligned}
n^{-1}\sum_{i=1}^{n} X_i &= \mu_1'(\boldsymbol{\theta}),\\
n^{-1}\sum_{i=1}^{n} X_i^2 &= \mu_2'(\boldsymbol{\theta}),\\
&\vdots\\
n^{-1}\sum_{i=1}^{n} X_i^n &= \mu_k'(\boldsymbol{\theta})
\end{aligned}
$$

という同時方程式が得られる（$\boldsymbol{\theta} = (\theta_1, \theta_2 \cdots, \theta_k)$）ので、これらを各 θ に関して、解けば、推定量（$\hat{\boldsymbol{\theta}} = (\hat{\theta_1}, \hat{\theta_2}, \cdots \hat{\theta_k})$）が得られる。これらをモーメント推定量と呼ぶ。

上式の左辺は任意の k 次のモーメントであるため、もし k 次のモーメントがわかるのであれば（例えば、k=1 なら標本平均、k=2 なら標本分散が既知であれば）、その式から代数的に母集団のパラメーターを推定した方が便利である。

テキスト例題 6.6：問題省略

解答

$E[X_1] = \mu$, $E[X_1^2] = \sigma^2 + \mu^2$ より、モーメント推定量は、上で示したモーメント推定量に関する方程式を用いると、$\bar{X}$=μ, $n^{-1}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2$ の解となる。

$$
\begin{aligned}
\hat{\mu} &= \bar{X},\\
\hat{\sigma^2} &= n^{-1}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\\
&= n^{-1}\sum_{i=1}^{n} X_i^2 - 2\bar{X}^2 + \bar{X}^2\\
&= n^{-1}\sum_{i=1}^{n} X_i^2 - 2n^{-1}\sum_{i=1}^{n} X_i \cdot \bar{X} + \bar{X}^2\\
&= n^{-1}(\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i \cdot \bar{X} + \bar{X}^2)\\
&= n^{-1}\sum_{i=1}^{n} (X_i - \bar{X})^2
\end{aligned}
$$

2.2 最尤法

そのデータが生起する尤もらしさを尤度といい、その尤度を表す関数、すなわち尤度関数は直感的に記述できる。今パラメーター θ の下、データ $\boldsymbol{X}$ がサンプリングされたとすると、その尤度関数は、

$$L(\boldsymbol{\theta} \mid \boldsymbol{X}) = \prod_{i=1}^{n} f(X \mid \theta) \tag{4}$$

と表される。最尤法の目的は、パラメーター θ を推定することなので、上式の尤もらしさを表す関数である、尤度関数を最大化するような θ を求めれば良いことになる。そして尤度関数を最大にするときの θ を最尤推定量 ($\hat{\theta}$) といい、数学的な定義で表記すれば、θ と $\hat{\theta}$ には以下のような関係が成り立つ。

$$L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}) = \sup_{\theta} L(\boldsymbol{\theta} \mid \boldsymbol{X}) \tag{5}$$

実際に最尤推定量を求める際には尤度方程式と呼ばれる、尤度関数を θ に関して微分した式=0 でおいた方程式を解けば、$\hat{\theta}$ が求まる。さらに、因子分解定理を適用すれば、得られた $\hat{\theta}$ が十分統計量であることも導ける。

最尤推定量の不変性

最尤法では、尤度関数に対して、対数変換を施した対数尤度関数を考える場合がある。対数尤度関数から最尤推定量を用いる際には、元のパラメーターに関する最尤推定量だけでなく、オッズやロジットといった統計量に関する最尤推定量が議論になることもある。そこで問題になるのが母数パラメータ θ に対して、何かしらの変換処理を施した変数も最尤推定量として扱える。これが「最尤推定量の不変性」と呼ばれる性質である。つまり、元のパラメーター θ に関する最尤推定量が求まれば、ロジットやオッズなどの統計量に関しても、最尤推定量が求まるということである。以下では「最尤推定量の不変性」の証明を述べる。

Proof. 変数変換を施す関数を $g(\cdot)$ とすると、$\tau = g(\theta)$ の推定をする際の、τ の最尤推定量が以下の式を満たすことを示せば良い。

$$L'(\hat{\boldsymbol{\tau}} \mid \boldsymbol{X}) = L'(\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \mid \boldsymbol{X}) \tag{6}$$

今、$\tau = g(\theta)$ という変数変換を考えているので、式 (5) で示される最尤推定量の定義式にならって、尤度関数 L'($\boldsymbol{\tau} \mid \boldsymbol{X}$) を以下の式で定義する。

$$L'(\boldsymbol{\tau} \mid \boldsymbol{X}) = \sup_{\{\theta \mid g(\theta)=\tau\}} L(\boldsymbol{\theta} \mid \boldsymbol{X}) \tag{7}$$

τ に関する上界をであることを考慮すれば、最尤推定量の定義より、

$$\sup_{\tau} L'(\boldsymbol{\tau} \mid \boldsymbol{X}) = \sup_{\theta} L(\boldsymbol{\theta} \mid \boldsymbol{X}) \tag{8}$$

が成り立つ。式 (7) と式 (8) より、τ の最尤推定量は

$$L'(\hat{\boldsymbol{\tau}} \mid \boldsymbol{X}) = \sup_{\tau} \sup_{\{\theta \mid g(\theta)=\tau\}} L(\boldsymbol{\theta} \mid \boldsymbol{X}) \tag{9}$$

と表される。上式の右辺は、($g(\boldsymbol{\theta})$=$\boldsymbol{\tau}$ という変数変換が施される場合の $\boldsymbol{\theta}$ に関する $L(\boldsymbol{\theta} \mid \boldsymbol{X})$ の上界) の $\boldsymbol{\tau}$ に関する上界である。つまり $\boldsymbol{\tau}$ に関する上界と言いつつ、$\boldsymbol{\theta}$ に関する上界を考えれば良いことになる。

よって式 (9) は

$$\begin{aligned} L'(\hat{\boldsymbol{\tau}} \mid \boldsymbol{X}) &= \sup_{\tau}\sup_{\{\theta|g(\theta)=\tau\}} L(\boldsymbol{\theta} \mid \boldsymbol{X}) \\ &= \sup_{\theta} L(\boldsymbol{\theta} \mid \boldsymbol{X}) \\ &= L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}) \end{aligned}$$

という、元の変数 $\boldsymbol{\theta}$ の最尤推定量 $\hat{\boldsymbol{\theta}}$ が満たす式に帰着できる。最後の等式は式 (5) を用いている。$\hat{\boldsymbol{\theta}}$に関する上界なので、g($\boldsymbol{\theta}$)=$g(\hat{\boldsymbol{\theta}})$ の場合の $\boldsymbol{\theta}$ という条件の元での、$\boldsymbol{\theta}$ に関する上界とも捉えられるので、

$$L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}) = \sup_{\{\theta|g(\theta)=g(\hat{\theta})\}} L(\boldsymbol{\theta} \mid \boldsymbol{X}) \tag{10}$$

となる。さらに式 (7) において今は、$\tau = g(\theta)$ が $\tau' = g(\hat{\theta})$ と対応していると考えれば、変数変換器 $g(\cdot)$ の中身 θ が $\hat{\theta}$ の場合の、θ に関する上界を考えていることになるので、

$$L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}) = \sup_{\{\theta|g(\theta)=g(\hat{\theta})\}} L(\boldsymbol{\theta} \mid \boldsymbol{X}) = L'(g(\hat{\boldsymbol{\theta}}) \mid \boldsymbol{X}) \tag{11}$$

となる。式 (7) と式 (11) より

$$\sup_{\{\theta|g(\theta)=g(\hat{\theta})\}} L(\boldsymbol{\theta} \mid \boldsymbol{X}) = L'(g(\hat{\boldsymbol{\theta}}) \mid \boldsymbol{X}) \tag{12}$$

式 (6) と式 (12) より、

$$L'(\hat{\boldsymbol{\tau}} \mid \boldsymbol{X}) = L'(g(\hat{\boldsymbol{\theta}}) \mid \boldsymbol{X}) \tag{13}$$

となるので、$\boldsymbol{\tau} = g(\boldsymbol{\theta})$ であることを考慮すると、$g(\hat{\boldsymbol{\theta}})$ は $g(\boldsymbol{\theta})$ の MLE となる。 □

2.3 ベイズ法

母集団のパラメーター $\boldsymbol{\theta}$ については「真の値を特定することはできないが限りなく近い値なら、推定することが可能」という前提でここまで議論してしてきたが、「我々が知り得ない母集団のパラメーターの真の値は1つに決まるものなのだろうか？」という問題提起が、以下で述べる「ベイズ法」を用いた推定の入り口である。

上で提起した問題に沿って話を進めるのであれば、推定しようとしているパラメーター $\boldsymbol{\theta}$ がある確率分布に従う、つまり $\boldsymbol{\theta}$ を確率変数とみなして議論を進めれば良い。$\boldsymbol{\theta}$ が従う確率分布は、その確率分布のパラメーターを ξ とすれば、$\pi(\boldsymbol{\theta} \mid \xi)$ と表すことができる。（詳細は後述するが）パラメーターを ξ は推定モデルの作成者が任意に決めれるパラメーター（ハイパーパラメーター）であるので、表記簡略のため、以下では、$\pi(\boldsymbol{\theta} \mid \xi) = \pi(\boldsymbol{\theta})$ として議論を進める。

サンプリングデータの集合を $\boldsymbol{X} = \boldsymbol{x}$ の場合に、このデータを表現する確率モデルは、確率変数 $\boldsymbol{X}$ と $\boldsymbol{\theta}$ が独立であることを考慮すれば、条件付き確率の定義より、

$$\begin{aligned} f(\boldsymbol{x}, \theta) &= f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \pi(\boldsymbol{\theta} \mid \boldsymbol{x}) f(\boldsymbol{x}) \end{aligned}$$

となるので、

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{x})} \tag{14}$$

となる。式 (14) の意味は、推定したいパラメーター $\boldsymbol{\theta}$ の確率分布に対して、パラメーター θ の場合の $\boldsymbol{x}$ が生起する確率（尤度）を通すことによって、より観測データを反映させたパラメーターの分布 $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ が得られることを意味しており、$\boldsymbol{X} = \boldsymbol{x}$ として**データが観測された後の θ の条件付き分布**なので、これを θ の**事後分布**といい、反対にデータ $\boldsymbol{X} = \boldsymbol{x}$ が得られる前の条件付き確率分布を**事前分布**と言う。

上で述べた、パラメーターの事後分布から統計量を推定量を推定する手法がベイズ法であり、ベイズ法を用いて算出される推定量を**ベイズ推定量**と言う（事後分布を算出するだけでなく、その事後分布の期待値などを算出することにより推定する)。また、詳細は割愛するが、$\boldsymbol{\theta}$ の事後分布も因子分解定理を満たすので、「十分統計量」であることを前提に、事後分布からベイズ推定量を算出すれば良い。ベイズ法の問題点として、事前分布の決め方がある。つまり、ハイパーパラメーター ξ の決め方である。このパラメーターが（データを手にしている）モデル作成者の主観により決められてしまうパラメータでもあるので、事前分布はしばしば「主観確率」とも呼ばれる。詳細は 6.4 章で言及されているが、「経験ベイズ法」や「階層ベイズ法」と言った手法は、この主観性を低減させるため（恣意性を除外するため）のハイパーパラメーター ξ の決め方である。

参考文献

[1] 鈴木武 (2014) 「数理統計学」 内田老鶴圃

[2] 須山敦志 (2017) 「機械学習スタートアップシリーズ ベイズ推論による機械学習入門 (KS 情報科学専門書)」講談社

Appendix

よく知られている 2 項分布の確率質量関数からは「指数型分布に属する」という結果は想像し難い。そこで 2 項分布の指数型関数の導出過程を以下に示す。

$$\begin{aligned}
f(x \mid p) &= {}_n\mathrm{C}_x p^x (1-p)^{n-x} \\
&= {}_n\mathrm{C}_x (1-p)^n \frac{p^x}{(1-p)^x} \\
&= {}_n\mathrm{C}_x (1-p)^n \exp\left\{ \log \frac{p^x}{(1-p)^x} \right\} \\
&= {}_n\mathrm{C}_x (1-p)^n \exp\left\{ x \log \frac{p}{(1-p)} \right\}
\end{aligned}$$