ヒストグラムに基づく
ナイーブなモード推定量の性質について

Properties of the naive mode estimator based on histograms

数学専攻　浅田 拓哉
ASADA, Takuya

1　はじめに

単峰の確率密度関数をもつ連続型確率変数に対し, モード (最頻値) はその最大確率密度をもつ変数の値として一意に定義される. モードは分布を代表する値であり, ベイズ推論などにおいても重要な意味を持つ.

モードを推定する方法として様々なものがある. その 1 つに, 初等中等教育でも扱われるヒストグラムを用いる方法 (ナイーブモード推定量) がある. これは, ヒストグラムを作成した後に度数が最大となる階級の階級値を母集団モードの推定値とするものであるが, 階級のとりかたによってヒストグラムの様子が変わってしまうため, 推定値に影響が出てしまう. また, その推定量としての良さは明らかではない.

それ以外の推定量として, 例えば Bickel and Fruhwirth (2006) では Half Sample Mode (HSM) を提案した. 他には Anderews and Bickel (1972) は Shorth, Rousseuw (1987) は location of the Least Median of Squares (LMS) を提案した. これらの推定量は一致性や計算の速度などの観点から評価がなされている.

そこで, 本論文では, ヒストグラムによる分布推定とナイーブモード推定量との関係を概観する. 次に, その他の方法の例として HSM, Shorth および LMS を概説する. 最後に, ナイーブモード推定量の推定量としてのよさを検証するため, 数値比較をおこなう.

2　ヒストグラムによる分布推定とモード

ヒストグラムは母集団分布をノンパラメトリックに推定する方法と考えることができる. その際, 階級の数あるいは階級の幅といったチューニングパラメータの設定が必要となる. ここでは, ヒストグラムとモードの推定の関係について扱い, 続いてヒストグラムによる分布推定と, チューニングパラメータの設定法について概観する.

2.1　ヒストグラムとモード

上述したように, ヒストグラムにおいて最大度数をもつ階級の階級値を標本モードとする方法は階級のとりかたに依存し, さらにその推定量としての良さは必ずしも明らかではない. ヒストグラムの用途は主に分布の形を把握することであり, 階級幅や階級数の決定は恣意的である. しかし, ヒストグラムに基づく標本モードの性質および一致性を議論する際には, サンプルサイズ n の増加に伴って階級数を大きくする必要がある. これにより, ヒストグラムは真の確率分布に近づいていくことから, 標本モードは一致性を有することが予想できる. サンンプルサイズ n に依存した階級数や階級幅の設定方法には様々な方法が提案されており, ここではその中から Sturges の方法 (Sturges, 1926), Scott の方法 (Scott, 1979) を紹介する. 他の方法については修士論文を参照されたい.

2.2 Sturges の方法

階級数を与える方法として広く知られたものが Sturges の方法 (Sturges, 1926) である．サンプルサイズを n として, Sturges の方法による階級数 $\hat{m}$ は

$$\hat{m} = 1 + \log_2 n$$

で与えられる．

Sturges の方法は正規分布と二項分布の関係に基づいて導出された方法である．正規分布からの大きさ n の標本が得られたとする．二項分布の正規近似 (ドモアブル・ラプラスの定理) より, $m+1$ 個の階級に分けたとき, 各階級の度数が二項分布 $\mathrm{Bin}(m, \frac{1}{2})$ における確率におおよそ比例していると考えられる．そこで, 二項係数の総和が標本の大きさ n に等しいとおくと,

$$n = \sum_{i=0}^{m} \binom{m}{i} = \sum_{i=0}^{m} \binom{m}{i} \cdot 1^i \cdot 1^{m-i} = 2^m$$

となり, $m = \log_2 n$ が導ける．ここで階級数 $\hat{m}$ は $m+1$ 個であるので, 先ほどの式を代入すれば Sturges の公式が得られる．

2.3 Scott の方法

Sturges の方法と同様に広く知られたものとして Scott の方法 (Scott, 1979) がある．Scott の方法は, 階級数ではなく階級幅 h を

$$h = \left[\frac{6}{n \int_{-\infty}^{\infty} \{f'(x)\}^2 dx} \right]^{\frac{1}{3}}$$

で与える方法である．

Scott の方法は, 真の確率密度関数とヒストグラムの確率密度関数の差を最小にする方法である．各階級 $\{B_k\}$ が

$$B_k = [t_k, t_{k+1}) \quad (k = -\infty, \ldots, -1, 0, 1, \ldots, \infty, \quad t_0 = 0)$$

という半開区間で作られているとする．ここで $\{t_k\}$ は $t_k < t_{k+1}$ を満たす格子点としている．また階級 B_k での度数を v_k, 等間隔な階級幅を h とすると, 階級 B_k での確率密度関数 $\hat{f}(x)$ は

$$\hat{f}(x) = \frac{v_k}{nh}$$

となる．この式は階級 B_k での相対度数を階級幅で割ったものである．ここで真の確率密度関数を $f(x)$ として, 積分平均二乗誤差 (Integraled Mean Squared Error, IMSE) は

$$\mathrm{IMSE} = \mathrm{E}\left[\int_{-\infty}^{\infty} \left(\hat{f}(x) - f(x) \right)^2 dx \right] = \int_{-\infty}^{\infty} \mathrm{E}\left[\left(\hat{f}(x) - f(x) \right)^2 \right] dx$$

となる．この IMSE を漸近的に最小にする階級幅が Scott の公式である．また, 修士論文にて, ナイーブモード推定量のバイアス, 分散, MSE を調べた結果, Scott の方法によるナイーブモード推定量がどの項目においても最小となった．

3 モードの推定

前節では, ヒストグラムにおいて最大度数をもつ階級の階級値とする方法, ヒストグラムの階級数や階級幅を決定する方法, およびカーネル密度推定とそのモード推定量の性質について説明した. しかしながら, これらの方法はモードを推定するための最良の方法というわけではなく, 十分な数学的性質も明らかになっていない. 本節では, ヒストグラムやカーネル密度推定を経由せず, 標本から直接モードを推定する方法について述べる.

3.1 Half Sample Mode

Bickel and Fruhwirth (2006) は, モードがデータの最も密集している値であると考え, 次のような推定法を提案した.

いま, $x_1, x_2, \ldots, x_n$ を連続型分布から得られた無作為標本に基づく順序統計量とする. すなわち $x_1 \leq x_2 \leq \cdots \leq x_n$ である. このとき, モード θ の推定量 $\hat{\theta}$ を以下のようにして与えるのが Half sample mode (HSM) である:

(1) $\frac{n}{2}$ 個が含まれる区間 $[x_{j_1}, x_{j_1+\frac{n}{2}-1}]$ の幅が最小となる j_1 を見つける. 言い換えると, $x_{j_1+\frac{n}{2}-1} - x_{j_1}$ が最小となる j_1 を求める. ただし, n が奇数の場合は $n+1$ を代わりに用いて計算する. $(1 \leq j_1 \leq \frac{n}{2}+1)$.
(2) 次に (1) で求めた区間内で, $\frac{n}{4}$ 個が含まれる区間 $[x_{j_2}, x_{j_2+\frac{n}{4}-1}]$ の幅が最小となる j_2 を求める. ここでも $\frac{n}{2}$ が奇数ならば $\frac{n}{2}+1$ を用いて計算する. $(j_1 \leq j_2 \leq \frac{n}{4}+j_1)$.
(3) ここまでの手順を 2 点 x_{j_k}, x_{j_k+1} になるまで続ける.
(4) 最後に得られた 2 点の平均, つまり以下の式がモードの推定量である:

$$\hat{\theta} = \frac{x_{j_k} + x_{j_k+1}}{2}.$$

このようにして, データが最も密集している区間を見つけようとしているのが HSM である. また, この推定方法はロバスト性を有しており, 計算速度の速い推定量の 1 つである.

3.2 Shorth と location of Least Median of Squares

Anderews and Bickel (1972) , Rousseeuw (1987) らは次のような推定方法を提案した.

HSM と同様に $x_1, x_2, \ldots, x_n$ を連続型分布から得られた無作為標本に基づく順序統計量とする. このとき, Shorth によるモード θ の推定量 $\hat{\theta}_{\mathrm{Shorth}}$, location of the Least Median of Squares (LMS) によるモード θ の推定量 $\hat{\theta}_{\mathrm{LMS}}$ は以下のようにして与えられる:

(1) 整数 h を, n が偶数のとき $h = \frac{n}{2}$, n が奇数のとき $h = \frac{n-1}{2}$ とする. このとき, 区間 $[x_k, x_{k+h}]$ の幅が最小となる k を求める.
(2) (1) で求めた区間 $[x_k, x_{k+h}]$ において Shorth と LMS の推定量は以下で与えられる:

$$\hat{\theta}_{\mathrm{Shorth}} = \frac{1}{h+1}\sum_{i=0}^{h} x_{k+i}, \qquad \hat{\theta}_{\mathrm{LMS}} = \frac{x_{k+h} - x_k}{2}.$$

ここで紹介した推定量は, Shorth は標本の約半数が密集している区間の重心を表しており, LMS は区間の中点の位置を表している. Shorth と LMS は, HSM と同様に, 外れ値の影響を受けないというロバスト性を有し

ており, 計算速度の速い推定量の 1 つである. また, 約半数が密集している区間を求めていることから, チューニングパラメータを設定しなくてよいことがわかる. 修士論文では, Shorth, LMS を一般化した方法 (Venter, 1967) についても, 理論的な性質も含めて紹介している.

4 まとめ

本論文では, ナイーブモード推定量の性質について数値実験を交えながらまとめた. 概して言えば, ナイーブモード推定量は一致性を有していることが伺えた. しかし, 初等中等教育で扱われる区切りのよい階級設定の場合は, 分散が不安定になるケースがあり, モード推定の観点からは適切ではないと考えられる. 一方, ヒストグラムによらないモード推定量とバイアス, 分散, MSE が大差ないことから, 計算の容易さの観点からナイーブモード推定量を用いて問題ないと考えられる. だが, これは特定の階級幅におけるナイーブモード推定量で考えたため, 一般の場合には注意が必要である.

課題点としては, ナイーブモード推定量の不偏性の有無およびヒストグラムの端点の変化によるナイーブモード推定量への影響が考えられる.

参考文献

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R.,Huber, P.H., Roger, W.H., and Tukey, J.W. (1972). *Robust Estimates of Location*, Princeton University Press.

[2] Bickel, D.R and Fruhwirth, R. (2006). On a fast, robust estimator of the mode: comparisons to other robust estimators with applications. *Computational Statistics and Data Analysis*, Vol.50, pp.3500-3530.

[3] Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association*,Vol.79, pp.871-880.

[4] Scott, D.W. (1979). On optimal and data-based histograms. *Biometrika*, Vol.66, No.3, pp.605-610.

[5] Sturges, H.A. (1926). The choice of a class interval. *Journal of the American Statistical Association*. Vol. 21, No.153 pp.65-66.

[6] Venter, J.H. (1967). On Estimation of the Mode. The Annals of Mathematical Statistics, Vol.38, No.5, pp.1446-1455.