

ヒストグラムの最適ビン幅の補正について

金沢大学大学院 人間社会環境研究科 齊藤実祥

金沢大学 人間社会研究域 寒河江雅彦

1. はじめに

与えられた定義域のもとで、ヒストグラム推定を考える。ヒストグラムの始点を定義域の左端点に定め、最適ビン幅で再構成された推定区間は一般に定義域と一致しない。以降、この定義域と推定区間のずれを「ビン残差」と呼ぶ。このビン残差を解消する補正法として、ビン残差を各ビンに等分配して調整することを考える。本報告では、この補正法に関して大標本における理論的性質を導き、有限サンプルについての数値実験を行い、ビン幅補正の有効性を示す。

2. 補正後ビン幅の理論的性質

定義域 $[a, b]$ が与えられ、 a と b は固定されたものとする。サンプル数 n 、最適ビン幅 h^* 、この時のビン数は $m^* = \lfloor \frac{b-a}{h^*} + \frac{1}{2} \rfloor$ で与えられる。ビン残差 δ は確率密度 $g(x)$ に従い、 $\delta \in \left[-\frac{h^*}{2}, \frac{h^*}{2}\right]$ とする。ここで、補正後ビン幅を $\tilde{h} = h^* - \frac{\delta}{m^*}$ とする。

補正後ビン幅のヒストグラム推定に関して次の大標本特性を証明した。

定理 1 漸近一致性

ビン幅 h について、 $n \rightarrow \infty$ のとき $h \rightarrow 0$ かつ $nh \rightarrow \infty$ であり、関数 $f(x)$ は絶対連続関数で、一階微分可能である時、

$$\hat{f}(x; \tilde{h}) \xrightarrow{d} f(x),$$

が MISE(平均積分二乗誤差) の意味で漸近的に成り立つ。

定理 2 漸近正規性

$h \propto O(n^{-\alpha})$, $x \in B_k$ に対して、

$$\alpha = \frac{1}{3} \text{ のとき、 } \sqrt{nh^*} \left\{ \hat{f}(x; \tilde{h}) - f(x) \right\} \xrightarrow{d} N \left(\text{Bias}[\hat{f}(x; \tilde{h})], f(\xi_k) \left(1 + \frac{\delta}{m^* h^*} \right) \right),$$

$$\alpha > \frac{1}{3} \text{ のとき、 } \sqrt{nh^*} \left\{ \hat{f}(x; \tilde{h}) - f(x) \right\} \xrightarrow{d} N \left(0, f(\xi_k) \left(1 + \frac{\delta}{m^* h^*} \right) \right),$$

が漸近的に成り立つ。ただし、 $\xi_k \in B_k$ は平均値の定理 $p_k = \int_{B_k} f(t)dt = hf(\xi_k)$ を満たす点で、 $\frac{\delta}{m^* h^*} \sim O(n^{-\frac{1}{3}})$ である。

ビン幅補正後 MISE のバウンド

補正後ビン幅を用いたヒストグラムの漸近的 MISE(AMISE) のバウンドは、

$$\text{AMISE}[\hat{f}(x; h^*)] \leq E_\delta \left[\text{AMISE}[\hat{f}(x; \tilde{h}) | h^*] \right] \leq \text{AMISE}[\hat{f}(x; h^*)] + \frac{R(f')}{48m^{*2}} h^{*2},$$

ただし、 $R(f') = \int f'(x)^2 dx$ である。

有限サンプルでの性質については数値実験に基づいて調べた。ビン残差を解消するためのビン幅補正によって、ヒストグラム推定の誤差と推定値の分散が低減していることから、その有効性を示した。

理論的性質および数値実験結果の詳細については当日報告する。

参考文献

Scott, D.W. (1979), On Optimal and Data-Based Histograms, *Biometrika*, Vol.66, pp.605-610.

Turuta, Y. and Sagae, M. (2017), Higher Order Kernel Density Estimation on the Circle, *Statistics & Probability Letters*, Vol.131, pp.46-50.