

研究タイトル：データの「かたち」を見える化する 統計手法の研究



氏名：	山本 けい子／YAMAMOTO Keiko	E-mail：	kei_yama@hakodate-ct.ac.jp
職名：	教授	学位：	博士(工学)
所属学会・協会：	日本統計学会, 日本計算機統計学会		
キーワード：	ノンパラメトリック統計解析, データ可視化, 確率密度関数の推定		
技術相談 提供可能技術：	・データ分布の可視化 ・統計的データ解析 ・ノンパラメトリック密度推定		

研究内容：

本研究では、データ分布の可視化および確率密度推定に関する研究として、ヒストグラムとカーネル密度推定(Kernel Density Estimation: KDE)の特性を踏まえた新たな推定手法の開発に取り組んでいる。ヒストグラムは計算効率に優れる一方で、ビン幅やアンカーポジション(ビン開始位置)の設定によって分布形状の見え方が大きく変化するという課題を有する。一方、KDE は滑らかな密度推定が可能であるものの、計算量が多く、帯域幅の設定に結果が大きく依存することが知られている。

そこで本研究では、ヒストグラムの計算効率と KDE の推定性能を両立する手法として、一般化 Frequency Polygon (Generalized Frequency Polygon: GFP)を提案した。GFP は従来の Frequency Polygon を拡張した確率密度推定法であり、少ない計算量で高い推定精度を実現できることを特徴としている。さらに、推定量の理論的性質を解析し、推定誤差や漸近特性を明らかにした。

図 1 は、同一データに対して異なるビン設定で作成したヒストグラムの例である。ヒストグラムはビン幅やビン位置によって分布構造の見え方が大きく変じ、場合によっては本来の特徴を適切に表現できないことがある。また、Excel, R, Python などのソフトウェアによってデフォルト設定が異なるため、利用時にはその特性を理解することが重要である。

図 2 は、二峰型分布に対して Frequency Polygon 系手法および KDE を適用した結果を示している。赤線で示した真の密度関数と比較すると、提案した GFP は少ない計算量でありながら良好な推定結果を与え、分布構造を適切に捉えていることが分かる。

さらに、本研究では教育データ、アンケートデータ、IoT センサデータなどの実データへの適用を進めている。分布の偏り、多峰性、外れ値の影響を視覚的かつ定量的に評価することで、従来の手法では見落とされやすかったデータの特徴を効率的に抽出する分析支援基盤の構築を目指している。

本研究の成果は、データサイエンスをはじめ、教育工学、IoT データ解析、品質管理、異常検知など、多様な分野におけるデータ分析支援への応用が期待される。

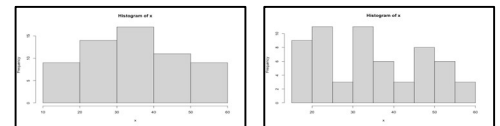


図 1. 同じデータに対するヒストグラム

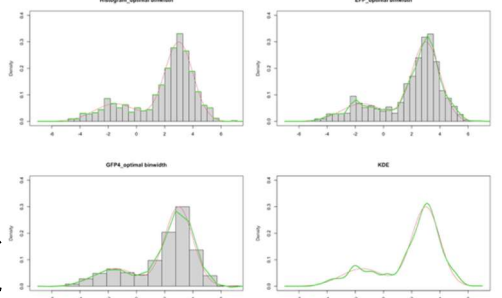


図 2. FP 型推定と KDE の比較

提供可能な設備・機器：

名称・型番(メーカー)	