ビン型データ構造を持つノンパラメトリック確率密度関数の推定に関する研究

齊藤 実祥

令和4年 12月

博士学位論文

ビン型データ構造を持つノンパラメトリック確率密度関数の推定に関する研究

金沢大学大学院 人間社会環境研究科 人間社会環境学専攻

学籍番号　1921082005

氏名　齊藤 実祥

主任指導教員名　寒河江 雅彦

目次

1　はじめに

1.1　研究背景と目的

データ構造を統計モデルを用いて推定する場合、大規模データと小規模データでそれぞれ異なる課題がある。大規模データ分析では、ビッグデータの活用が様々な分野で期待されるものの、ビッグデータの 3 つの特徴である Volume(量)、Velocity(速度)、Variety(多様性) に対して、メモリー容量や計算速度等のハードウェアとソフトウェアの性能上の制約、計算アルゴリズムにおける計算量の増加で、従来の統計モデルのままでは対応できなくなっている。それに対して、小規模データ分析においても、データ量の少なさによる情報不足や推定結果の不安定さ等で、データの構造を正確に捉えることが難しいという問題がある。そこで、これらの大規模データ及び小規模データ解析の問題点に対処した統計分析手法について研究する。

統計分析手法は、パラメトリック統計手法とノンパラメトリック統計手法の 2 つに大別されるが、本稿ではノンパラメトリック統計手法に焦点を当てる。ノンパラメトリック統計手法には 2 つのメリットが挙げられる。1 つ目は、データに柔軟に対応が可能で、モデルに柔軟性がある。複雑な構造を取り扱いやすく、例えば、同一のモデルで単峰から二山以上の多峰の分布まで分析が可能である。2 つ目は、あらかじめ特定の母集団分布を仮定せずに推定が可能であり、母集団構造が特定できない状況において相性が良い方法と考えられる。本稿では特に、データ構造を規定する確率密度推定に焦点を当て、その中でもデータ縮約化ができるビン型の密度推定法について議論する。ビン型の密度推定法の中で最も簡単でよく知られている Histogram 推定法に焦点を絞る。Histogram 推定法について、以降では簡単のため「Histogram」と略す。

Histogram は、データを重複しない区間に分け、各区間に入るデータ数に比例した柱状グラフを構成して密度推定を行う方法である。この重複しない区間は「ビン」と呼ばれ、ビン幅及び推定区間が Histogram を同定するためのパラメータとみなすことができる。Histogram ではビンごとに独立に区分的定数関数で推定することでデータを縮約できるが、元データを復元することは不可能であるため、重要な情報まで破棄している場合がある。そこで、この情報損失を回避するために大規模データと小規模データのそれぞれにおける Histogram の改良を考える。

大規模データ解析における情報技術的な処理の方向には、データ圧縮、ハードウェアとソフトウェアの組み合わせ、計算アルゴリズム等がある。これらに加えて、統計モデルの改良で大規模データに対応するアプローチとして平滑化やモーメント等の付加情報の利用による Histogram の改良法がある。従来、Histogram の拡張法については、2 つの研究方向がある。一つは、各ビンにおける平均や分散等の局所モーメント情報を利用する方法である。もう一つは、各ビンの境界で不連続性を解消するために用いられる平滑化法である。先行研究については 2.4 節で詳述するが、いずれも Histogram より推定精度が改良されることが示されている。しかしながら、局所モーメント情報の利用と Histogram の平滑化を同時に行う推定モデルについては未整備である。したがって、本研究では局所モーメント情報の利用と平滑化を同時に満たす多項式型の平滑化 Polynomial

Histogram 推定量を提案する。その理論的性質の導出及び数値実験から、従来型のビン型推定法に比べて推定精度が改良できることを示す。

小規模データにおける情報不足による推定の不安定化を回避するため、推定後のビン幅補正を行う方法とデータを繰り返し利用した改良法の 2 つの方法について検討する。

前者は、Histogram が小規模データにおいて、ビン幅が広く推定されやすいことから生じる推定区間の問題に対する改良法である。閉区間の定義域が与えられている場合に、最適ビン幅に基づいて構築した Histogram の推定区間と定義域は一致しない。Histogram のビン幅が広い場合、この推定区間と定義域とのずれが大きく、ずれの範囲内のデータは推定に使用されないため、無視されるデータが多くなる。そこで、推定区間と定義域が一致するようにビン幅を補正する方法を提案する。その理論的性質の導出と数値実験からビン幅補正の有効性について明らかにする。

後者は、Histogram の最適ビン幅推定において、未知の母集団分布 f をその推定量 $\hat{f}$ で置き換える必要があるが、データを繰り返し利用することによって $\hat{f}$ の推定を回避する改良法である。具体的には、データのリサンプリングと反復計算を組み合わせた推定法であるブートストラップ法から着想を得た分布構造に依存しない distribution-free なビン幅決定法を考える。分割点を一様乱数で決定した不等間隔 Histogram を複数回推定し、その平均を推定量とする Random Partitioned Histogram を提案する。提案モデルのパラメータに関して複数パターンの数値実験を行い、有限標本において Histogram を改良できることを示す。

1.2 本稿の構成

本稿の構成は図 1.1 の通りである。研究全体の流れとしては大規模データ解析のための Histogram の改良と、小規模データ解析のための Histogram の改良の 2 つに大別される。本稿は、2 章で Histogram の改良に関する先行研究について整理し、3 章と 4 章で大規模データ解析のための Histogram の改良について、5 章と 6 章で小規模データ解析のための Histogram の改良について述べる。具体的な構成は次の通りである。

2.1 と 2.2 節で Histogram の定義及び理論的性質を説明する。2.3 節で、Histogram のビン数とビン幅推定の改良に関わる先行研究について述べる。2.4 節で、Histogram を拡張する既存の推定モデルについて、主に漸近的性質の観点から整理する。

3 章と 4 章で、局所モーメント情報の条件と 2 次連続性の条件を同時に満たす平滑化 Polynomial Histogram を提案し、その理論的性質と数値実験から、既存の主なビン型推定量と比較して推定精度が改良されることを示す。

5 章で、既知の定義域と Histogram の推定区間が一致するようにビン幅を補正する方法を提案し、その理論的性質と数値実験から提案手法が実用的に有効であることを示す。6 章で、ランダムな分割点による不等間隔 Histogram を複数回推定し、その平均を推定量とする Random Partitioned Histogram を提案し、有限標本において Histogram を改良することを示す。

最後に 7 章で、本稿における結論と考察及び本研究の今後の展開方向について議論する。

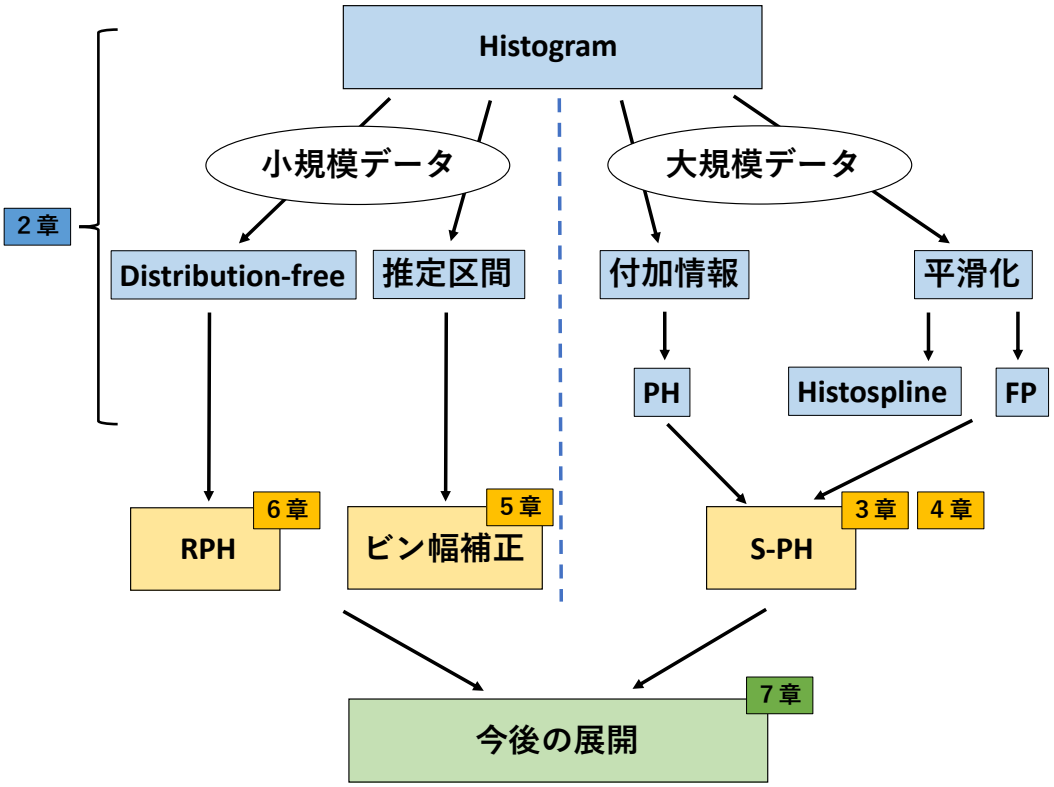

図 1.1　本稿の構成 (青色：先行研究、黄：本研究、緑色：今後の展開)

2　ビン型確率密度関数

本章では、Histogram の定義及び漸近的性質と、ビン数とビン幅及び拡張モデルに関する先行研究について述べる。

2.1　Histogram の定義

確率密度関数 $f(x)$ からの n 個の標本 $\{X_1, X_2, \ldots, X_n\}$、$j$ 番目のビン B_j、B_j の範囲 $[y_{j-1},\ y_j)$、ビン幅 h、B_j に入る度数 ν_j とする。このとき、B_j における Histogram は次式で与えられる;

$$\hat{f}_{HIST}(x) = \frac{\nu_j}{nh} = \frac{1}{nh}\sum_{i=1}^{n} I_{[y_{j-1},\ y_j)}(X_i)\ , \quad x \in B_j, \tag{2.1}$$

ただし、$I_A(x)$ は定義関数で次の通りである;

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A. \end{cases}$$

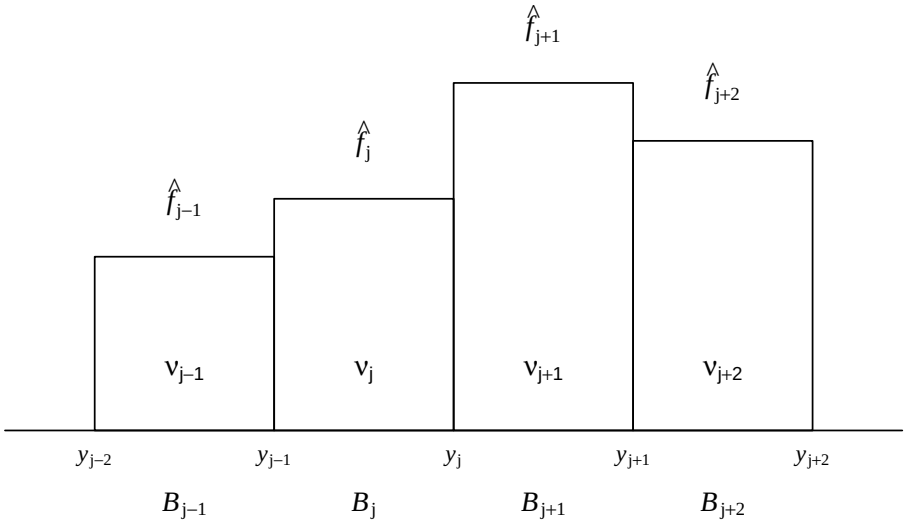

図 2.1　Histogram の構造

2.2 Histogram の漸近的性質

推定モデルの良さを測る二乗誤差基準として、推定量 $\hat{f}$ と真の密度関数 f との平均積分二乗誤差 (以降、MISE) がある。MISE の定義は次の通りである;

$$\begin{aligned}\mathrm{MISE} := E\left[\int\left\{\hat{f}(x)-f(x)\right\}^2 dx\right] &= \int E\left[\hat{f}(x)-f(x)\right]^2 dx \\ &= \mathrm{IV}\left\{\hat{f}(x)\right\} + \mathrm{ISB}\left\{\hat{f}(x)\right\}, \end{aligned} \tag{2.2}$$

ただし、IV と ISB は次の通り定義される;

$$\begin{aligned}\mathrm{IV}\left\{\hat{f}(x)\right\} &= \int \mathrm{Var}\left\{\hat{f}(x)\right\} dx, \\ \mathrm{ISB}\left\{\hat{f}(x)\right\} &= \int \mathrm{Bias}\left\{\hat{f}(x)\right\}^2 dx,\end{aligned}$$

ここで、分散及びバイアスは、以下の通りである;

$$\begin{aligned}\mathrm{Var}\left\{\hat{f}(x)\right\} &= E\left[\left(\hat{f}(x)-E\left[\hat{f}(x)\right]\right)^2\right], \\ \mathrm{Bias}\left\{\hat{f}(x)\right\} &= E\left[\hat{f}(x)\right]-f(x).\end{aligned}$$

以降、推定量の漸近的性質の評価は MISE 基準に基づいている。

Histogram の漸近的な MISE(以降、AMISE) は、

$$\begin{aligned}\mathrm{AMISE}[\hat{f}_{HIST}(x)] &= \mathrm{AIV}_{HIST} + \mathrm{AISB}_{HIST} \\ &= \frac{1}{nh} + \frac{R(f')}{12}h^2, \end{aligned} \tag{2.3}$$

ただし、

$$R(\phi) = \int \phi(x)^2 dx,$$

である。

このとき、AMISE を最小化する意味で最適なビン幅 h^* と最小 AMISE_{HIST} は、

$$h^* = \left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}, \tag{2.4}$$

$$\text{最小 } \mathrm{AMISE}_{HIST} = \frac{3}{2}\left(\frac{R(f')}{6}\right)^{\frac{1}{3}} n^{-\frac{2}{3}}. \tag{2.5}$$

$R(f')$ は真の分布 f に依存する未知の定数とみなすことができるため、推定精度は標本数 n に依存する。したがって、Histogram の漸近的な推定精度は $O\left(n^{-\frac{2}{3}}\right)$ となる。

2.3 Histogram のビン数・ビン幅推定

2.3 節では、これまでに提案された Histogram のビン数とビン幅推定法で主要なものについて述べる。

2.3.1 スタージェスのルール

2.3.1 と 2.3.2 項ではビン数の推定法に関する議論について述べる。

Histogram のビン数決定に関するルールで古典的なものは、Sturges(1926) が提案した手法である。Sturges は正規曲線を二項分布 $B(m,\ p=0.5)$ で近似することで Histogram を作成した。この近似による標本は、二項係数 $\binom{m}{j},\ j=0,1,\ldots,m$ によって与えられるビンの度数 ν_j である。この二項係数との関係から、$m+1$ 個のビンを持つ Histogram が与えられる。この Histogram は単峰で左右対称の形であり、両端は 0 に収束する。

度数の総和、すなわち標本数 n は、

$$n=\sum_{j=0}^{m}\nu_j=\sum_{j=0}^{m}\binom{m}{j}=\sum_{j=0}^{m}\binom{m}{j}1^j1^{m-j}=2^m,$$

となる。m について解くと、

$$m=\log_2 n,$$

が得られる。このときのビン数を $K_{Sturges}$ とすると、$K_{Sturges}=m+1$ より、

$$K_{Sturges}=1+\log_2 n,$$

となり、上式がスタージェスのルールである。

2.3.2 ドアネのルール

ビンの数に関するもう一つの有名なルールには、Doane(1976) が提案したものが挙げられる。前述した通り、Sturges は正規曲線を二項分布で近似していることから、スタージェスのルールは基本的に対称かつ単峰な分布を仮定したもとでのビン数推定のルールである。データが歪んでいたり、尖りが大きい場合には追加のビンが必要になる可能性がある。そこで Doane はスタージェスのルールに歪度の情報を加えて、これをビン数とすることを提案した。

この時のビン数を K_{Doane} とすると、

$$K_{Doane}=1+\log_2 n+k_e,$$

ここで $k_e=\log_2\left(1+\frac{\sqrt{b1}}{\sigma_{\sqrt{b1}}}\right)$ であり、$\sqrt{b1}$ は歪度の統計量、$\bar{x}$ は標本平均、$\sigma_{\sqrt{b1}}$ は $\sqrt{b1}$ の標準偏差で、次の通りである;

$$\sqrt{b1}=\frac{|\sum(x-\bar{x})^3|}{|\sum(x-\bar{x})^2|^{\frac{3}{2}}},\quad \sigma_{\sqrt{b1}}=\sqrt{\frac{6(n-2)}{(n+1)(n+3)}}.$$

K_{Doane} は歪度が 0 の場合に、スタージェスのルールと等しくなる。

2.3.3　スコットのルール

2.3.3 項以降は、ビン幅の推定に関する議論について述べる。

Scott(1979) は二乗誤差基準の一つである MISE を用いて理論的にビン幅推定を行う手法を提案した。(2.4) の通り、最適なビン幅の推定を行うためには $R(f')$ を解く必要がある。しかしながら、密度関数 f は未知であることから、当然 f' も未知であり、$R(f')$ は自明でない値である。そのため、Scott は参照分布として $f = N(\mu,\ \sigma^2)$ を用いることを考えた。この時の $R(f')$ は、

$$R(f') = \frac{1}{4\sqrt{\pi}\sigma^3},$$

であるため、

$$h^* = \left(\frac{24\sqrt{\pi}\sigma^3}{n}\right)^{\frac{1}{3}} \sim 3.5\sigma n^{-\frac{1}{3}},$$

となる。ただし、標準偏差 σ は未知のスケールパラメータであるため、標本標準偏差 $\hat{\sigma}$ で置き換えて、

$$\hat{h}_{Scott} = 3.5\hat{\sigma}n^{-\frac{1}{3}},$$

をスコットのルールとした。

2.3.4　フリードマン・ダイアコニスのルール

ビン幅推定に関するルールの 2 つ目に、フリードマン・ダイアコニスのルールが挙げられる。Freedman-Diaconis(1981) はよりスコットのルールより頑健性のあるルールとして、標準偏差 σ を四分位範囲 (IQR) の倍数で置き換えて、

$$\hat{h}_{FD} = 2(\mathrm{IQR})n^{-\frac{1}{3}},$$

をビン幅とすることを提案した。データが正規分布に従う場合には、IQR=1.348σ から、フリードマン・ダイアコニスのルールはスコットのルールの約 77% となり、より狭いビン幅を推定する。

2.3.5　テレル・スコットのルール

Terrell and Scott(1985) は、既知の区間 $[a,\ b]$ についてスコットのルールによる最適なビン幅 $\hat{h}_{Scott}$ 時のビン数 K^* を求め、ビン数の下限を導出した。

$$K^* = \frac{b-a}{\hat{h}_{Scott}} \geq \frac{2}{\left(\frac{4}{n}\right)^{\frac{1}{3}}} = \sqrt[3]{2n}.$$

2.3.6　クロス・バリデーション法

クロス・バリデーション法は、カーネル型密度関数の最適バンド幅を探すための方法である。Rudemo(1982)、Bowman(1984) によって提案された最小二乗クロス・バリデーション (LSCV)

を Histogram の推定に応用することで、最適なビン幅が選択される。LSCV での最適なビン幅は、

$$\mathrm{LSCV}(h) = \int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i),$$

ただし、$\hat{f}_{-i}(x_i)$ は次の通りである;

$$\hat{f}_{-i}(x_i) = \frac{\nu_j - 1}{(n-1)h}, \quad x_i \in B_j.$$

2.3.7 プラグイン法

プラグイン法についても MISE の最小化に基づく方法であるとみなすことができる。プラグイン法は、最適なバンド幅を見つけるために反復解法を必要とし、これまで述べてきた推定問題の中で最も複雑なものとなる。

プラグイン法によって得られるビン幅 $\hat{h}_{PI}$ について、

$$\hat{h}_{PI} = \left\{ \frac{6}{-\hat{\psi}_2(g)n} \right\}^{\frac{1}{3}},$$

からの未知の項 $R(f')$ をカーネル推定量 $\hat{\psi}_2$ で置き換える。すなわち、

$$\psi_2 = E[f^{(2)}(x)] = \int_{-\infty}^{\infty} f^{(2)}(x) f(x) dx,$$

である。L が十分滑らかで対称な単峰型カーネル関数とすると、$\hat{\psi}_2$ の推定量は、

$$\psi_2(g) = n^{-2} g^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n} L^{(2)} \left\{ \frac{(X_i - X_j)}{g} \right\},$$

となる。ただし、g はカーネル L のバンド幅である。

2.4 Histogram の拡張

2.4 節では、これまでに提案されてきた Histogram の拡張法について、主に理論的性質の側面から述べる。

2.4.1 Polynomial Histogram

Polynomial Histogram(以降、PH) は、各ビンにおける平均や分散等の局所モーメント情報を利用する推定法である。Sagae and Scott(1997) と Scott and Sagae(1997) は高次までの局所モーメントを追加的に利用した PH 推定法を提案し、MISE の意味で高次オーダーの収束性を持つことを示している。

局所モーメントの次数を q とすると、各ビン $B_j = [y_{j-1},\ y_j)$ における q 次 PH 推定量は、未知のパラメータ $\alpha_0, \alpha_1, \ldots, \alpha_q$ を用いて、以下の通り定義される;

$$\hat{f}_q(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_q x^q, \quad x \in B_j.$$

標本数 n、ビン B_j、度数を ν_j、ビンの中点 t_j、l 次の局所モーメント $\mu_j^{(l)},\ (l = 0, 1, \ldots, q)$ とする。このとき、局所モーメント情報を保持するように以下の制約条件に基づいて $\alpha_0, \alpha_1, \ldots, \alpha_q$ を求めることで q 次 PH 推定量が得られる;

$$\int_{B_j} \hat{f}_q(x)dx = \frac{\nu_j}{n} := \mu_j^{(0)},$$
$$\int_{B_j} (x - t_j)^l \hat{f}_q(x)dx \Big/ \int_{B_j} \hat{f}_q(x)dx = \mu_j^{(l)}.$$

q 次 PH 推定量 $\hat{f}_q(x)$ の MISE 基準に基づく推定精度は、$O\left(n^{-\frac{2q+2}{2q+3}}\right)$ であり、$q = 0$ の時 Histogram に帰着し、$q \geq 1$ で改良される。

2.4.2 Frequency Polygon

Scott(1985) によって、Frequency Polygon(以降、FP) が提案された。FP は各ビンの中点を節点として Histogram を線形補間することで、各ビンの境界の不連続性を解消する方法である。

節点数 N、ビンの端点 $y_j,\ (j = 0, 1, \ldots, N)$、ビンの中点 t_j、$x \in [t_j,\ t_{j+1}]$ における FP 推定量 $\hat{f}_{FP}(x)$ は次の通り定義される;

$$\hat{f}_{FP}(x) = \left(1 - \frac{x - t_j}{h}\right) \hat{f}_j(x) + \left(\frac{x - t_j}{h}\right) \hat{f}_{j+1}(x),$$

ただし、$\hat{f}_j(x)$ はビン B_j における Histogram 推定量である。

FP の AMISE$[\hat{f}_{FP}(x)]$、最適ビン幅 h^*_{FP}、最小 AMISE$_{FP}$ は次の通りである;

$$\text{AMISE}[\hat{f}_{FP}(x)] = \frac{2}{3nh} + \frac{49}{2880} h^4 R\left(f''\right),$$
$$h^*_{FP} = 2\left(\frac{15}{49R\left(f''\right)}\right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$
$$\text{最小 AMISE}_{FP} = \frac{5}{12}\left(\frac{49R\left(f''\right)}{15}\right)^{\frac{1}{5}} n^{-\frac{4}{5}}.$$

FP の MISE に基づく推定精度は $O\left(n^{-\frac{4}{5}}\right)$ で、Histogram の $O\left(n^{-\frac{2}{3}}\right)$ を改良している。

2.4.3 Edge Frequency Polygon

Jones et al.(1998) によって Edge Frequency Polygon(以降、EFP) が提案された。EFP は、各ビンの端点を節点とし、隣接ビンの高さの平均を節点の高さとして Histogram を線形補間する方法である。$x \in [y_{j-1}, y_j)$ における EFP 推定量 $\hat{f}_{EFP}(x)$ は次の通り定義される;

$$\hat{f}_{EFP}(x) = \left(\frac{1}{2} - \frac{x - t_j}{h}\right) \frac{\hat{f}_{j-1}(x) + \hat{f}_j(x)}{2} + \left(\frac{1}{2} + \frac{x - t_j}{h}\right) \frac{\hat{f}_j(x) + \hat{f}_{j+1}(x)}{2}.$$

EFP の AMISE$[\hat{f}_{EFP}(x)]$、最適ビン幅 h^*_{EFP}、最小 AMISE_{EFP} は次の通りである;

$$\text{AMISE}[\hat{f}_{EFP}(x)] = \frac{5}{12nh} + \frac{23}{360}h^4 R\left(f''\right),$$
$$h^*_{EFP} = \left(\frac{75}{46R\left(f''\right)}\right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$
$$\text{最小 AMISE}_{EFP} = \frac{25}{48}\left(\frac{46R\left(f''\right)}{75}\right)^{\frac{1}{5}} n^{-\frac{4}{5}}.$$

最小 AMISE の定数は FP が 0.5279、EFP が 0.4723 であり、EFP は FP を改良していることが分かる。

2.4.4 Bias-Optimized Frequency Polygon

Minnotte(1996, 1998) によって Bias-Optimized Frequency Polygon(以降、BFP) が提案された。FP と同様に各ビンの中点を節点とし、節点の高さは各ビンで面積相等性を満たすように選択して Histogram を非線形補間する方法である。累積分布関数の連続性の次数を p とすると、$x \in [t_j,\ t_{j+1}]$ における $\text{BFP}_{(p-1)}$ 推定量 $\hat{f}_{BFP(p-1)}(x)$ は次の通り定義される;

$$\hat{f}_{BFP(p-1)}(x) = \frac{1}{nh}\sum_{\kappa=1}^{p-1}\sum_{i} b_{i\kappa}\nu_{j+1}\left(\frac{x-t_j}{h}\right)^{\kappa},$$

ただし、κ は多項式における次数、$b_{i\kappa}$ は i 次ベルヌーイ多項式 Be_i を用いて定義される多項式で、p は 2 以上の正の偶数である。

$\text{BFP}_{(p-1)}$ の AMISE$[\hat{f}_{BFP(p-1)}(x)]$ は次の通りである;

$$\text{AMISE}[\hat{f}_{BFP(p-1)}(x)] = \frac{C_1(p)}{nh} + C_2(p)h^{2p}\int f^{(p)}(x)^2 dx,$$

ただし、$C_1(p),\ C_2(p)$ は次の通りである;

$$C_1(p) = \sum_i \sum_{\kappa_1=0}^{p-1}\sum_{\kappa_2=0}^{p-1} \frac{b_{i\kappa_1}b_{i\kappa_2}}{(\kappa_1+\kappa_2+1)},$$
$$C_2(p) = -\frac{1}{(2p)!}Be_{2p}.$$

$\text{BFP}_{(p-1)}$ の推定精度は $O\left(n^{-\frac{2p}{2p+1}}\right)$ である。

2.4.5 Histospline

Boneva, Kendall and Stefanov(BKS)(1971) と Schoenberg(1973) によって BKS 型 Histospline が提案され、Lii and Rosenblatt(LR)(1974) が LR 型 Histospline の MISE に基づく推定精度を導出した。齊藤・寒河江 (2021) は BKS 型 Histospline と LR 型 Histospline の理論的同等性を示し、MISE の陽な漸近表現を導いている。以降、BKS 型 Histospline 及び LR 型 Histospline を簡略化のため Histospline と呼ぶ。

Histospline は、ビンの端点を節点とし、面積相等性を満たしながら、各ビンの境界で累積分布関数の 2 次連続性まで満たすように Histogram を滑らかな曲線で接続する方法である。ビン B_j 上の Histospline 推定量 $\hat{f}_{HS}(x)$ は次の通り定義される;

$$\hat{f}_{HS}(x) = \frac{\mu_j^{(0)}}{h} + \left\{ -\frac{1}{24h} + \frac{1}{2h^2}(x-t_j) + \frac{1}{2h^3}(x-t_j)^2 \right\} \sum_{k=1}^{N-1} w_{j,k} \left(\mu_{k+1}^{(0)} - \mu_k^{(0)} \right) + \left\{ \frac{1}{24h} + \frac{1}{2h^2}(x-t_j) - \frac{1}{2h^3}(x-t_j)^2 \right\} \sum_{k=1}^{N-1} w_{j-1,k} \left(\mu_{k+1}^{(0)} - \mu_k^{(0)} \right),$$

ただし、$\mu_j^{(0)}$ はビン B_j の面積、$w_{j,k} = \frac{3}{\sqrt{3}} \left(\sqrt{3} - 2 \right)^{|j-k|}$ である。

Histospline の AMISE$[\hat{f}_{HS}(x)]$、最適ビン幅 h_{HS}^*、最小 AMISE$_{HS}$ は次の通りである;

$$\text{AMISE}[\hat{f}_{HS}(x)] = \left(\frac{5\sqrt{3}+3}{10} \right) \frac{1}{nh} + \frac{R\left(f'''\right)}{30240} h^6,$$

$$h_{HS}^* = \left(\frac{2520\sqrt{3}+1512}{R\left(f'''\right)} \right)^{\frac{1}{7}} n^{-\frac{1}{7}},$$

$$\text{最小 AMISE}_{HS} = \frac{35\sqrt{3}+21}{60} \left(\frac{R\left(f'''\right)}{2520\sqrt{3}+1512} \right)^{\frac{1}{7}} n^{-\frac{6}{7}}.$$

Histospline の MISE に基づく推定精度は $O\left(n^{-\frac{6}{7}}\right)$ で、各ビンの境界における連続性を満たすことで Histogram を改良できる。

2.4.6 Averaged Shifted Histogram

Scott(1985) によって Averaged Shifted Histogram(以降、ASH) が提案された。ASH は Histogram を構築した後、同じデータに対して分割点をシフトさせた Histogram を再構築し、それらの平均を推定量とするモデルである。

同じデータセットに対してビン幅 h の Histogram を τ 回シフトさせて推定し、各回での Histogram の左端点は $\{ih/\tau,\ i = 0, \ldots, \tau - 1\}$ とする。メッシュに区切った後のビン幅 $\frac{h}{\tau}$、k 番目のメッシュ $I_k = \left[\frac{(k-1)h}{\tau},\ \frac{kh}{\tau} \right)$、$i$ 番目の Shifted Histogram 推定を $\hat{g}_i(x)$ とすると、$x \in I_k$ における ASH 推定量 $\hat{f}_{ASH}(x)$ は次の通りに定義される;

$$\hat{f}_{ASH}(x) = \frac{1}{\tau} \sum_{i=0}^{\tau-1} \hat{g}_i(x).$$

ASH の AMISE$[\hat{f}_{ASH}(x)]$ は次の通りである;

$$\text{AMISE}[\hat{f}_{ASH}(x)] = \frac{2}{3nh} \left(1 + \frac{1}{2\tau^2} \right) + \frac{h^2}{12\tau^2} R\left(f'\right) + \frac{h^4}{144} \left(1 - \frac{2}{\tau^2} + \frac{3}{5\tau^4} \right) R\left(f''\right).$$

シフト数 $\tau \to \infty$ において、ASH の MISE に基づく推定精度は $O\left(n^{-\frac{4}{5}}\right)$ であり、Histogram を改良している。

前述した Histogram を拡張する既存手法と本研究の推定法において満たす条件について整理すると表 2.1 の通りである。

表 2.1　先行研究と本研究の推定法における設定 (青：先行研究、黄：本研究、緑：今後の展開)

局所モーメント / 連続性	−	0	1	2	q
−	—	経験分布関数	—	—	—
0	—	Histogram PH(0)	PH(1)	PH(2)	PH(q)
1	FP, EFP ASH	BFP(1)			
2		BFP(2) Histospline	S-PH(2,1)	S-PH(2,2)	S-PH(2, q)
p		BFP(p)			S-PH(p, q)

2 章　3 章　4 章　7 章

3　平滑化 Polynomial Histogram：S-PH$_{(2,1)}$

本章では、1 次までの局所モーメント情報の利用と節点における累積分布関数の 2 次連続性を同時に満たす多項式 Polynomial Histogram(以降、S-PH) について論じる。

3.1　推定量の構築

提案モデルはモーメントの次数と連続性の次数に依存するため、S-PH$_{(p,q)}$ で表記する。ここで、p は累積分布関数の連続性の次数、q は局所モーメントの次数とする。これは、従来のビン型推定量を包括した表現であり、Histogram は S-PH$_{(0,0)}$、2 次 BFP は S-PH$_{(1,0)}$、1 次 PH は S-PH$_{(0,1)}$、Histospline は S-PH$_{(2,0)}$ に対応する。本章では累積分布関数の 2 次までの連続性条件と、局所 1 次までのモーメント条件を同時に満たす S-PH$_{(2,1)}$ の場合を扱う。

以降、簡単のために記法について説明する。標本数 n、節点数 N、ビン節点 $y_j,\ (j = 0, 1, \ldots, N)$、定義域 $[y_0,\ y_N]$ とする。j 番目のビン $B_j = [y_{j-1},\ y_j),\ (j = 1, 2, \ldots, N)$、ビン幅 h、ビンの中点 $t_j = \frac{y_{j-1}+y_j}{2}$、各ビンの面積 $\mu_j^{(0)}$、局所 1 次モーメント $\mu_j^{(1)}$、j 番目のビン B_j における S-PH$_{(2,1)}$ 推定量 $\hat{F}_j(x)$、y_j における分布関数の高さ $G_j = \hat{F}_j(y_j)$、$\hat{F}_j(x)$ の係数 $a_0^{(j)}, a_1^{(j)}, \ldots, a_4^{(j)}$ とし、以降、$a_0, a_1, \ldots, a_4$ と略す。

この時、S-PH$_{(2,1)}$ 推定量 $\hat{F}_j(x)$ は次の通りである;

$$\hat{F}_j(x) = a_0 + a_1(x - t_j) + a_2(x - t_j)^2 + a_3(x - t_j)^3 + a_4(x - t_j)^4, \quad x \in B_j. \tag{3.1}$$

ここで、局所モーメント条件と各ビンの境界での連続性の条件を満たすために、以下の制約条件を設ける;

(i) 局所 0 次モーメント情報 (面積相等性)：$\int_{B_j}(x - t_j)^0 \hat{F}_j'(x)dx = \mu_j^{(0)}$,
(ii) 局所 1 次モーメント情報：$\int_{B_j}(x - t_j)^1 \hat{F}_j'(x)dx = \mu_j^{(1)}$,
(iii) 累積分布関数の 1 次連続性：$\hat{F}'(y_j-) = \hat{F}'(y_j+)$,
(iv) 累積分布関数の 2 次連続性：$\hat{F}''(y_j-) = \hat{F}''(y_j+)$,

ただし、$\hat{F}^{(i)}(y_j-),\ (i = 0, 1, \ldots)$ は節点 y_j における $\hat{F}^{(i)}(x)$ の左方 i 次微分係数、$\hat{F}^{(i)}(y_j+)$ は右方 i 次微分係数である。上記の条件のみでは未知数に対して制約条件の数が 2 つ不足する。そのため、節点 y_j における $\hat{F}_j(x)$ の 1 次微係数 $M_j^{(1)},\ (j = 0, 1, \ldots, N)$ に対し、付加条件として端条件 $M_0^{(1)} = M_N^{(1)} = 0$ を仮定する。

各係数について制約条件に基づく方程式を解くと以下を得る;

$$a_0 = -\frac{15}{8h}\mu_j^{(1)} + \frac{h}{32}\left(M_j^{(1)} - M_{j-1}^{(1)}\right) + \frac{1}{2}\left(G_j + G_{j-1}\right), \tag{3.2}$$

$$a_1 = \frac{3}{2h}\mu_j^{(0)} - \frac{1}{4}\left(M_j^{(1)} + M_{j-1}^{(1)}\right), \tag{3.3}$$

$$a_2 = \frac{15}{h^3}\mu_j^{(1)} - \frac{3}{4h}\left(M_j^{(1)} - M_{j-1}^{(1)}\right), \tag{3.4}$$

$$a_3 = -\frac{2}{h^3}\mu_j^{(0)} + \frac{1}{h^2}\left(M_j^{(1)} + M_{j-1}^{(1)}\right), \tag{3.5}$$

$$a_4 = \frac{5}{2h^3}\left(M_j^{(1)} - M_{j-1}^{(1)}\right) - \frac{30}{h^5}\mu_j^{(1)}, \tag{3.6}$$

ここで、未知の定数 $M_j^{(1)},\ (j = 1, 2, \ldots, N-1)$ について制約条件から行列の形 $\boldsymbol{RM} = \boldsymbol{d}$ で以下の通り表わされる;

$$\begin{pmatrix} 2 & -\frac{1}{3} & & & & & \\ -\frac{1}{3} & 2 & -\frac{1}{3} & & & \mathbf{0} & \\ & \ddots & \ddots & \ddots & & & \\ & & -\frac{1}{3} & 2 & -\frac{1}{3} & & \\ & & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & & -\frac{1}{3} & 2 & -\frac{1}{3} \\ & & & & & -\frac{1}{3} & 2 \end{pmatrix} \begin{pmatrix} M_1^{(1)} \\ M_2^{(1)} \\ \vdots \\ M_j^{(1)} \\ \vdots \\ M_{N-2}^{(1)} \\ M_{N-1}^{(1)} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_j \\ \vdots \\ d_{N-2} \\ d_{N-1} \end{pmatrix}, \tag{3.7}$$

ただし、$\boldsymbol{M} = \left(M_1^{(1)}, \ldots, M_{N-1}^{(1)}\right)^\top$、$\boldsymbol{d}$ の要素は $d_j = \frac{2}{3h}\left(\mu_{j+1}^{(0)} + \mu_j^{(0)}\right) - \frac{20}{3h^2}\left(\mu_{j+1}^{(1)} - \mu_j^{(1)}\right)$, $(j = 1, 2, \ldots, N-1)$ である。節点数 N が十分大きく、$M_j^{(1)}$ は定義域の端点から十分に離れたビンでの要素とすると、

$$M_j^{(1)} = \frac{1}{2h}\sum_{k=1}^{N-1} w_{j,k}\left\{\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}, \tag{3.8}$$

ここで、$w_{j,k} = \frac{1}{\sqrt{2}}\left(3 - 2\sqrt{2}\right)^{|j-k|}$、$\sum_{k=1}^{N-1}\frac{1}{\sqrt{2}}\left(3 - 2\sqrt{2}\right)^{|j-k|} = 1$ である。

したがって、$\hat{f}_j(x) = \hat{F}_j'(x)$ とすると、S-PH$_{(2,1)}$ の密度推定量 $\hat{f}_j(x)$ は次式で与えられる;

$$\begin{aligned} \hat{f}_j(x) = &\left\{-\frac{6}{h^3}\left(x - t_j\right)^2 + \frac{3}{2h}\right\}\mu_j^{(0)} + \left\{-\frac{120}{h^5}\left(x - t_j\right)^3 + \frac{30}{h^3}\left(x - t_j\right)\right\}\mu_j^{(1)} \\ &+ \left\{\frac{10}{h^3}\left(x - t_j\right)^3 + \frac{3}{h^2}\left(x - t_j\right)^2 - \frac{3}{2h}\left(x - t_j\right) - \frac{1}{4}\right\} \\ &\qquad \times \frac{1}{2h}\sum_{k=1}^{N-1} w_{j,k}\left\{\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\} \\ &+ \left\{-\frac{10}{h^3}\left(x - t_j\right)^3 + \frac{3}{h^2}\left(x - t_j\right)^2 + \frac{3}{2h}\left(x - t_j\right) - \frac{1}{4}\right\} \\ &\qquad \times \frac{1}{2h}\sum_{k=1}^{N-1} w_{j-1,k}\left\{\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}. \end{aligned} \tag{3.9}$$

3.2 漸近的性質

本節ではビンごとに推定された $\hat{f}_j(x)$ で構成された S-PH$_{(2,1)}$ 密度推定量について、漸近的 MISE の導出と漸近正規性が成り立つことを示す。

S-PH$_{(2,1)}$ 密度推定量に関して次の正則条件を満たすものとする;

(i) ビン幅 h について、$n \to \infty$ のとき、$h \to 0$ かつ $nh \to \infty$,
(ii) 関数 $f^{(4)}(x)$ は絶対連続関数で、$R\left(f^{(5)}\right) = \int f^{(5)}(x)^2 dx < \infty$.

この (i) と (ii) の条件のもとで、以下の S-PH$_{(2,1)}$ 推定量の漸近的な MISE と漸近正規性を以下の定理及び系にまとめる。

Theorem 1：S-PH$_{(2,1)}$ の AMISE

S-PH$_{(2,1)}$ 密度推定量 $\hat{f}_{SPH(2,1)}(x)$ の漸近的な MISE(AMISE) は、

$$\begin{aligned} \mathrm{AMISE}[\hat{f}_{SPH(2,1)}(x)] &= \left(\frac{992+695\sqrt{2}}{840}\right)\frac{1}{nh} + \frac{h^8}{1209600}R\left(f^{(4)}\right) \\ &\fallingdotseq \frac{2.351}{nh} + \frac{h^8}{1209600}R\left(f^{(4)}\right). \end{aligned} \tag{3.10}$$

このときの最適ビン幅 $h^*_{SPH(2,1)}$ 及び最小 AMISE$_{SPH(2,1)}$ は、

$$h^*_{SPH(2,1)} = \left(\frac{125100\sqrt{2}+178560}{R(f^{(4)})}\right)^{\frac{1}{9}} n^{-\frac{1}{9}}, \tag{3.11}$$

$$\text{最小 AMISE}_{SPH(2,1)} = \left(\frac{2976+2085\sqrt{2}}{2240}\right)\left(\frac{R\left(f^{(4)}\right)}{178560+125100\sqrt{2}}\right)^{\frac{1}{9}} n^{-\frac{8}{9}}. \tag{3.12}$$

Theorem 2：S-PH$_{(2,1)}$ の漸近正規性

S-PH$_{(2,1)}$ 密度推定量について、

$$\frac{\sum_{s=1}^{n} E\left|Z_s - E\left[Z_s\right]\right|^3}{\sigma\left[\hat{f}_j(x)\right]^3} = \frac{O\left(\frac{1}{n^2h^2}\right)}{O\left(\frac{1}{n^{3/2}h^{3/2}}\right)} = O\left(\frac{1}{n^{1/2}h^{1/2}}\right) = o(1), \tag{3.13}$$

ただし、$\sigma\left[\hat{f}_j(x)\right]^2 = \mathrm{Var}\left\{\hat{f}_j(x)\right\}$ で、Z_s, $(s = 1, 2, \ldots, n)$ は $\hat{f}_j(x)$ に従う独立な確率変数である。したがって、リアプノフの条件を満たすことから、S-PH$_{(2,1)}$ 密度推定量は漸近正規性が成り立つ。

Corollary 1：各ビンにおける S-PH$_{(2,1)}$ の漸近正規性

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

$\alpha = \frac{1}{9}$ のとき、

$$\sqrt{nh}\left\{\hat{f}_j(x) - f(x)\right\} \xrightarrow{d} N\left(\mathrm{Bias}\left[\hat{f}_j(x)\right], \left(\frac{992 + 695\sqrt{2}}{840}\right) f\left(\xi_j\right)\right), \tag{3.14}$$

$\alpha > \frac{1}{9}$ のとき、

$$\sqrt{nh}\left\{\hat{f}_j(x) - f(x)\right\} \xrightarrow{d} N\left(o(1),\ \left(\frac{992 + 695\sqrt{2}}{840}\right) f\left(\xi_j\right)\right), \tag{3.15}$$

が漸近的に成り立つ。ただし、$f(\xi_j)$ は $p_j = \int_{B_j} f(t)dt = hf(\xi_j),\ \xi_j \in B_j$ を満たす B_j 内のある点とする。Theorem 1 の S-PH$_{(2,1)}$ の AMISE については 3.4 節の Appendix 1、Theorem 2 の漸近正規性の証明については 3.5 節の Appendix 2 で詳述する。

3.3 数値実験

3.3.1 ISE 数値実験

S-PH$_{(2,1)}$ の有限標本における推定精度を調べるため、積分二乗誤差 (以降、ISE) について数値実験を行う。定義域 $[-8,\ 8]$ の混合正規分布 $f(x)$ は $\frac{3}{5}N\left(-3, 2^2\right) + \frac{2}{5}N\left(4, 1^2\right)$ に従う確率密度関数とし、データ数を $n = 10^2,\ 10^3,\ 10^4,\ 10^5$ とする。ビン幅は理論上の最適ビン幅を用いて、各データ数における ISE の数値実験 1000 回の平均 (MISE) を算出する。同様の設定のもとでの Histogram、1 次 PH、Histospline と比較する。

数値実験に用いる各推定量の最適ビン幅 h^* は次の通りである。

表 3.1　各推定量の最適ビン幅

(a)Histogram	$h^*_{HIST} = \left(\frac{6}{R(f^{(1)})}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}$
(b)1 次 PH	$h^*_{PH(1)} = \left(\frac{360}{R(f^{(2)})}\right)^{\frac{1}{5}} n^{-\frac{1}{5}}$
(c)Histospline	$h^*_{HSP} = \left(\frac{2520\sqrt{3}+1512}{R(f^{(3)})}\right)^{\frac{1}{7}} n^{-\frac{1}{7}}$
(d)S-PH$_{(2,1)}$	$h^*_{SPH(2,1)} = \left(\frac{125100\sqrt{2}+178560}{R(f^{(4)})}\right)^{\frac{1}{9}} n^{-\frac{1}{9}}$

図 3.1 は $n = 10^3$ のときの Histogram、1 次 PH、Histospline と提案モデル S-PH$_{(2,1)}$ の数値実験結果の例で、実線が推定結果、破線が真の密度関数のグラフである。

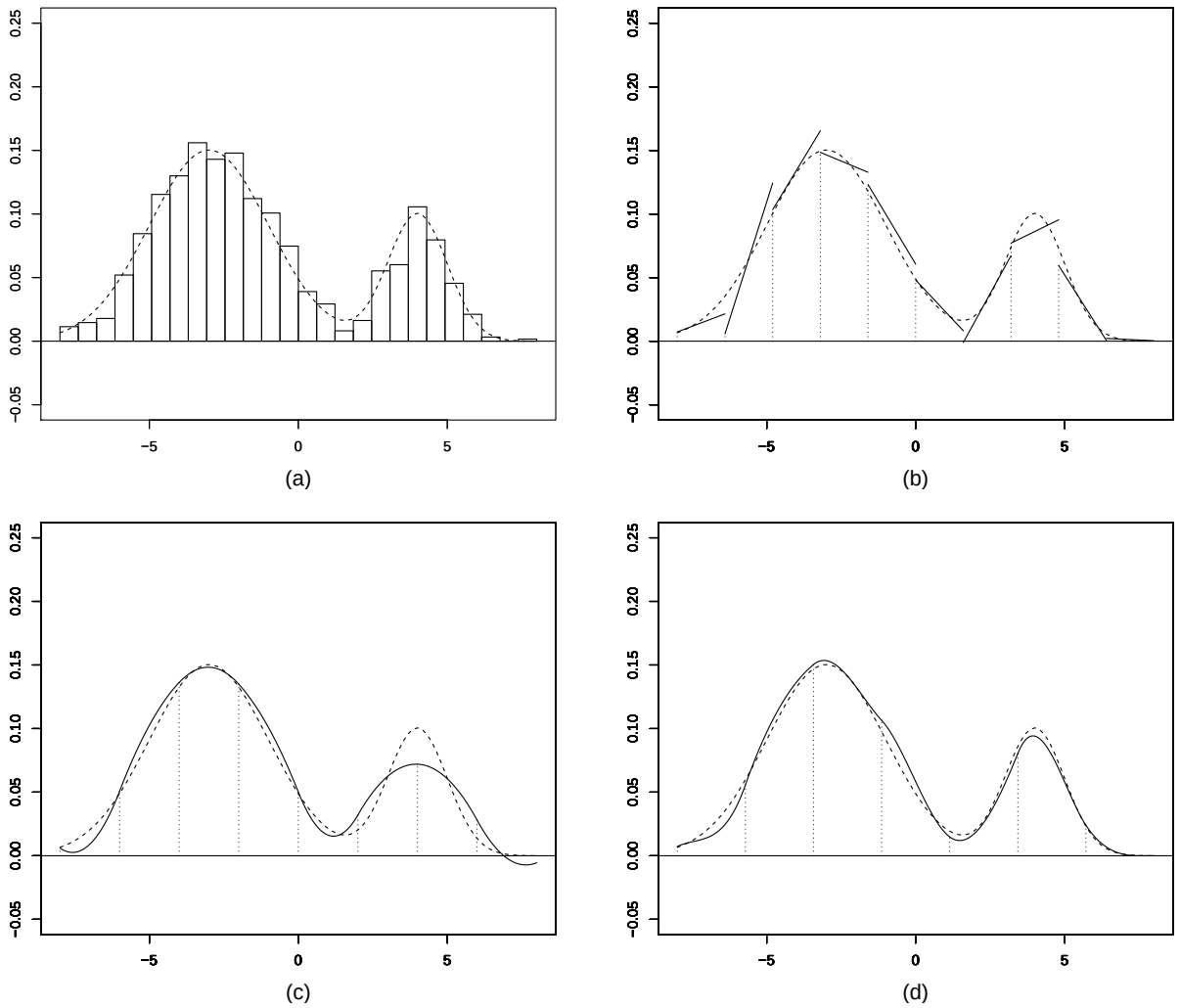

図 3.1　$n = 10^3$ での推定例 (実線が推定結果、破線が真の分布。(a)Histogram、(b)1 次 PH、(c)Histospline、(d)S-PH$_{(2,1)}$)

表 3.2　主な 3 つの推定量と S-PH$_{(2,1)}$ の MISE 比較 (繰り返し数 1000 回)

	MISE*	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
Histogram	$O\left(n^{-\frac{2}{3}}\right)$	0.013374	0.003088	0.000677	0.000147
PH$_{(1)}$	$O\left(n^{-\frac{4}{5}}\right)$	0.008047	0.001380	0.000230	0.000040
Histospline	$O\left(n^{-\frac{6}{7}}\right)$	<u>0.003893</u>	0.001756	0.000187	0.000021
S-PH$_{(2,1)}$	$O\left(n^{-\frac{8}{9}}\right)$	0.006218	<u>0.000849</u>	<u>0.000112</u>	<u>0.000016</u>

表 3.2 はデータ数 n を変化させた時、4 つの推定法を比較した MISE の実験結果である。n ごと

に比較して最小の値に下線を引いてある。どの推定量も標本数が大きくなるにつれて ISE は小さくなる。データ数が $n = 10^3$ 以上の時に、S-PH$_{(2,1)}$ の ISE が最も小さい。

表 3.3　Histogram との最適ビン幅比

	h^*	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
Histogram	$O\left(n^{-\frac{1}{3}}\right)$	1.0	1.0	1.0	1.0
PH$_{(1)}$	$O\left(n^{-\frac{1}{5}}\right)$	5.5	7.4	9.9	13.1
Histospline	$O\left(n^{-\frac{1}{7}}\right)$	7.3	12.0	18.8	27.9
S-PH$_{(2,1)}$	$O\left(n^{-\frac{1}{9}}\right)$	<u>11.0</u>	<u>16.0</u>	<u>25.9</u>	<u>44.6</u>

表 3.3 はデータ数 $n = 10^{2\sim5}$ と変化させた時の、Histogram の最適ビン幅でそれぞれの最適なビン幅を割った値である。つまり、Histogram のビン幅を 1.0 としたときの 1 次 PH、Histospline、S-PH$_{(2,1)}$ の最適なビン幅である。n を変化させた時の最大の値に下線を引いてある。データ数に関わらず S-PH$_{(2,1)}$ のビン幅が他の推定法と比較して最も大きい。例えば、$n = 10^5$ では Histogram の約 45 倍のビン幅である。推定したビン幅で定義域を割ることでビン数が得られるが、S-PH$_{(2,1)}$ は他の推定法と比較して最も少ないビン数で推定できる。

3.3.2　実データ解析

S-PH$_{(2,1)}$ 推定量を実データに適用し、提案手法の実用面での特徴を調べる。石川県「いしかわ・金沢　風と緑の楽都音楽祭 2019 実態分析報告」の石川県内消費額 $n = 1762$ 個のデータを用いて、元データに基づく標本平均と、Histogram、1 次 PH、Histospline、S-PH$_{(2,1)}$ の各推定量に基づく期待値を比較する。データは定義域 $[0, 300000]$ で、等間隔のビン 5 つで設定した。

表 3.4　元データ及び各推定量に基づく標本平均

	推定量に基づく標本平均
元データ	<u>19316.4</u>
Histogram	36061.3
1 次 PH	<u>19316.4</u>
Histospline	31616.2
S-PH$_{(2,1)}$	<u>19316.4</u>

表 3.4 はデータに基づく標本平均と、各推定量に基づく期待値である。1 次 PH と S-PH$_{(2,1)}$ では、付加情報として局所平均を利用しているため、推定量に基づく期待値が標本平均と一致してい

る。それに対し、Histogram と Histospline では局所平均の情報を保持していないため、元データに対して平均が過大評価されている。したがって、S-PH で局所モーメント情報を保持するように推定量を構築することで、ビン化に伴う情報損失の回避が可能になる。

3.4 Appendix 1：AMISE$\{\hat{f}(x)\}$ の証明

S-PH$_{(2,1)}$ 推定量の AMISE について漸近積分分散 AIV$\left\{\hat{f}(x)\right\}$ と漸近積分二乗バイアス AISB$\left\{\hat{f}(x)\right\}$ のそれぞれから導出する。

まず漸近積分二乗バイアスについて示す。S-PH$_{(2,1)}$ の密度推定量は (3.9) 式より、

$$\begin{aligned}\hat{f}_j(x) = &\left\{-\frac{6}{h^3}(x-t_j)^2+\frac{3}{2h}\right\}\mu_j^{(0)}+\left\{-\frac{120}{h^5}(x-t_j)^3+\frac{30}{h^3}(x-t_j)\right\}\mu_j^{(1)}\\ &+\left\{\frac{10}{h^3}(x-t_j)^3+\frac{3}{h^2}(x-t_j)^2-\frac{3}{2h}(x-t_j)-\frac{1}{4}\right\}\\ &\qquad\times\frac{1}{2h}\sum_{k=1}^{N-1}w_{j,k}\left\{\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}\\ &+\left\{-\frac{10}{h^3}(x-t_j)^3+\frac{3}{h^2}(x-t_j)^2+\frac{3}{2h}(x-t_j)-\frac{1}{4}\right\}\\ &\qquad\times\frac{1}{2h}\sum_{k=1}^{N-1}w_{j-1,k}\left\{\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}. \qquad (3.16)\end{aligned}$$

ここで (3.16) 式の期待値を取ると、

$$\begin{aligned}E\left[\hat{f}_j(x)\right] = &\, c_1E\left[\mu_j^{(0)}\right]+c_2E\left[\mu_j^{(1)}\right]\\ &+c_3\sum_{k=1}^{N-1}w_{j,k}\left\{E\left[\mu_{k+1}^{(0)}\right]+E\left[\mu_k^{(0)}\right]-\frac{10}{h}\left(E\left[\mu_{k+1}^{(1)}\right]-E\left[\mu_k^{(1)}\right]\right)\right\}\\ &+c_4\sum_{k=1}^{N-1}w_{j-1,k}\left\{E\left[\mu_{k+1}^{(0)}\right]+E\left[\mu_k^{(0)}\right]-\frac{10}{h}\left(E\left[\mu_{k+1}^{(1)}\right]-E\left[\mu_k^{(1)}\right]\right)\right\}, \qquad (3.17)\end{aligned}$$

ただし、

$$c_1 = -\frac{6}{h^3}(x-t_j)^2+\frac{3}{2h}, \qquad (3.18)$$

$$c_2 = -\frac{120}{h^5}(x-t_j)^3+\frac{30}{h^3}(x-t_j), \qquad (3.19)$$

$$c_3 = \frac{10}{h^3}(x-t_j)^3+\frac{3}{h^2}(x-t_j)^2-\frac{3}{2h}(x-t_j)-\frac{1}{4}, \qquad (3.20)$$

$$c_4 = -\frac{10}{h^3}(x-t_j)^3+\frac{3}{h^2}(x-t_j)^2+\frac{3}{2h}(x-t_j)-\frac{1}{4}. \qquad (3.21)$$

各ビンの度数 ν_k とし、ν_k は二項分布 $B(n,\ p_k)$, $(k=1,2,\ldots,N)$ に従い、$p_k = \int_{B_k} f(t)dt$

の時、

$$E\left[\mu_k^{(0)}\right] = \frac{1}{n}E\left[\nu_k\right] = \frac{1}{n}np_k = \int_{B_k} f(t)dt, \tag{3.22}$$

$$E\left[\mu_k^{(1)}\right] = \int_{B_k} (t-t_k)f(t)dt, \tag{3.23}$$

ここで、未知の $f(t)$ についてテイラー級数による近似から以下の通り表せられる;

$$\int_{B_k} f(t)dt \sim \int_{B_k} \Bigg\{ f(x) + (t-x)f^{(1)}(x) + \frac{1}{2!}(t-x)^2 f^{(2)}(x) + \frac{1}{3!}(t-x)^3 f^{(3)}(x) + \frac{1}{4!}(t-x)^4 f^{(4)}(x) + \frac{1}{5!}(t-x)^5 f^{(5)}(x) + \cdots \Bigg\} dt. \tag{3.24}$$

(3.17) 式で (3.18)〜(3.24) を用いて、$\sum_{k=1}^{N-1} w_{j,k}k = j,\ \sum_{k=1}^{N-1} w_{j,k}k^2 = j^2 + \frac{1}{2},$
$\sum_{k=1}^{N-1} w_{j,k}k^3 = j^3 + \frac{3}{2}j,\ \sum_{k=1}^{N-1} w_{j,k}k^4 = j^4 + 3j^2 + 2$ であることを利用して整理すると、
$\hat{f}_j(x)$ のバイアスは、

$$\begin{aligned} \mathrm{Bias}\left[\hat{f}_j(x)\right] &= E\left[\hat{f}_j(x)\right] - f(x) \\ &= -\frac{1}{24}f^{(4)}(x)\left\{(x-t_j)^4 - \frac{h^2}{2}(x-t_j)^2 + \frac{7}{240}h^4\right\}. \end{aligned} \tag{3.25}$$

したがって、ビン B_j における $\mathrm{AISB}\left\{\hat{f}_j(x)\right\}$ は、

$$\begin{aligned} \mathrm{AISB}\left\{\hat{f}_j(x)\right\} &= \int_{B_j} \frac{1}{576}f^{(4)}(\xi_j)^2 \Bigg\{(x-t_j)^8 - h^2(x-t_j)^6 + \frac{127}{120}(x-t_j)^4 - \frac{7}{240}h^6(x-t_j)^2 + \frac{49}{57600}h^8\Bigg\}dx \\ &= \frac{h^9}{1209600}f^{(4)}(\xi_j)^2. \end{aligned} \tag{3.26}$$

以上より、$\mathrm{AISB}\left\{\hat{f}(x)\right\}$ はリーマン積分近似 $\sum_j f^{(4)}(\xi_j)^2 h = \int f^{(4)}(x)^2 dx + o(1)$ を用いて、

$$\mathrm{AISB}\left\{\hat{f}(x)\right\} = \frac{R(f^{(4)})}{1209600}h^8. \tag{3.27}$$

次に漸近積分分散について示す。$\hat{f}_j(x)$ の分散は、

$$
\begin{aligned}
\mathrm{Var}\left[\hat{f}_j(x)\right] &= \mathrm{Var}\left(c_1\mu_j^{(0)}\right) + \mathrm{Var}\left(c_2\mu_j^{(1)}\right) \\
&\quad + \mathrm{Var}\left[c_3\left\{\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}\right] \\
&\quad + \mathrm{Var}\left[c_4\left\{\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}\right] \\
&\quad + 2\mathrm{Cov}\left(c_1\mu_j^{(0)},\ c_2\mu_j^{(1)}\right) \\
&\quad + 2\mathrm{Cov}\left(c_1\mu_j^{(0)},\ c_3\left\{\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}\right) \\
&\quad + 2\mathrm{Cov}\left(c_1\mu_j^{(0)},\ c_4\left\{\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}\right) \\
&\quad + 2\mathrm{Cov}\left(c_2\mu_j^{(1)},\ c_3\left\{\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}\right) \\
&\quad + 2\mathrm{Cov}\left(c_2\mu_j^{(1)},\ c_4\left\{\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}\right) \\
&\quad + 2\mathrm{Cov}\Bigg(c_3\left\{\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}, \\
&\qquad\qquad c_4\left\{\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)} + \mu_k^{(0)}\right) - \frac{10}{h}\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)} - \mu_k^{(1)}\right)\right\}\Bigg).
\end{aligned}
\tag{3.28}
$$

以降で、分散の評価では主要項のみ取り出して記述する。

(3.28) の第 1 項は、近似的に、

$$
\begin{aligned}
\mathrm{Var}\left(c_1\mu_j^{(0)}\right) = \frac{c_1^2}{n^2}\mathrm{Var}\left(\nu_j\right) &= \frac{c_1^2}{n^2} np_j(1-p_j) \\
&\sim \frac{c_1^2}{n} hf(x).
\end{aligned}
\tag{3.29}
$$

(3.28) の第 2 項は、

$$
\mathrm{Var}\left(c_2\mu_j^{(1)}\right) = c_2^2\mathrm{Var}\left(\mu_j^{(1)}\right) \sim c_2^2\frac{h^3}{12n} f(x). \tag{3.30}
$$

(3.28) の第 3 項は、

$$\begin{aligned}
&\mathrm{Var}\left[c_3\left\{\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}\right] \\
&\quad = c_3^2\left(1+\frac{100}{h^2}\right)\sum_{k=1}^{N-1} w_{j,k}^2\left\{\mathrm{Var}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)+\mathrm{Var}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\} \\
&\qquad + 2\mathrm{Cov}\left(c_3\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right), -\frac{10}{h}c_3\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right) \\
&\quad \sim \frac{31\sqrt{2}}{6}c_3^2\left(\frac{h}{n}\right)f(x). \qquad (3.31)
\end{aligned}$$

(3.28) の第 4 項は、(3.31) と同様に、

$$\begin{aligned}
&\mathrm{Var}\left[c_4\left\{\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}\right] \\
&\quad \sim \frac{31\sqrt{2}}{6}c_4^2\left(\frac{h}{n}\right)f(x). \qquad (3.32)
\end{aligned}$$

(3.28) の第 5 項は、

$$\begin{aligned}
2\mathrm{Cov}\left(c_1\mu_j^{(0)},\ c_2\mu_j^{(1)}\right) &= 2c_1c_2\mathrm{Cov}\left(\mu_j^{(0)},\mu_j^{(1)}\right) \\
&\sim c_1c_2\frac{h^3}{6n}f'(x). \qquad (3.33)
\end{aligned}$$

(3.28) の第 6, 7 項は、

$$\begin{aligned}
&2\mathrm{Cov}\left(c_1\mu_j^{(0)},\ c_3\left\{\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}\right) \\
&\quad + 2\mathrm{Cov}\left(c_1\mu_j^{(0)},\ c_4\left\{\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}\right) \\
&\sim 2c_1c_3\left\{\mathrm{Cov}\left(\mu_j^{(0)},\sum_{k=1}^{N-1} w_{j,k}\mu_{k+1}^{(0)}\right)+\mathrm{Cov}\left(\mu_j^{(0)},\sum_{k=1}^{N-1} w_{j,k}\mu_k^{(0)}\right)\right\} \\
&\quad + 2c_1c_4\left\{\mathrm{Cov}\left(\mu_j^{(0)},\sum_{k=1}^{N-1} w_{j-1,k}\mu_{k+1}^{(0)}\right)+\mathrm{Cov}\left(\mu_j^{(0)},\sum_{k=1}^{N-1} w_{j-1,k}\mu_k^{(0)}\right)\right\} \\
&\sim 2c_1(c_3+c_4)\frac{1}{n^2}\left(w_{j,j-1}+w_{j,j}+w_{j-1,j-1}+w_{j-1,j}\right)\mathrm{Var}(\nu_j) \\
&= 2\sqrt{2}\left(2-\sqrt{2}\right)c_1(c_3+c_4)\left(\frac{h}{n}\right)f(x). \qquad (3.34)
\end{aligned}$$

(3.28) の第 8, 9 項は、

$$
\begin{aligned}
&2\mathrm{Cov}\left(c_2\mu_j^{(1)},\ c_3\left\{\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}\right)\\
&\quad+2\mathrm{Cov}\left(c_2\mu_j^{(1)},\ c_4\left\{\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}\right)\\
&\sim-\frac{20}{h}c_2c_3\left\{\mathrm{Cov}\left(\mu_j^{(1)},\sum_{k=1}^{N-1} w_{j,k}\mu_{k+1}^{(1)}\right)-\mathrm{Cov}\left(\mu_j^{(1)},\sum_{k=1}^{N-1} w_{j,k}\mu_k^{(1)}\right)\right\}\\
&\quad-\frac{20}{h}c_2c_4\left\{\mathrm{Cov}\left(\mu_j^{(1)},\sum_{k=1}^{N-1} w_{j-1,k}\mu_{k+1}^{(1)}\right)-\mathrm{Cov}\left(\mu_j^{(1)},\sum_{k=1}^{N-1} w_{j-1,k}\mu_k^{(1)}\right)\right\}\\
&\sim-\frac{20}{h}c_2(c_3+c_4)\left(w_{j,j-1}-w_{j,j}+w_{j-1,j-1}-w_{j-1,j}\right)\mathrm{Var}\left(\mu_j^{(1)}\right)\\
&=\frac{5\sqrt{2}}{3}\left(1-\sqrt{2}\right)c_2(c_4-c_3)\left(\frac{h^2}{n}\right)f(x).
\end{aligned}
\tag{3.35}
$$

(3.28) の第 10 項は、$\sum_{k=1}^{N-1} w_{j,k}w_{j,k+1}=\frac{\sqrt{2}}{8}$, $\sum_{k=1}^{N-2} w_{j,k}w_{j-1,k+1}=\frac{64-45\sqrt{2}}{8}$, $\sum_{k=1}^{N-2} w_{j-1,k}^2=\frac{3\sqrt{2}}{8}$ を用いて、

$$
\begin{aligned}
&2\mathrm{Cov}\Bigg(c_3\left\{\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\},\\
&\qquad\qquad c_4\left\{\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)}+\mu_k^{(0)}\right)-\frac{10}{h}\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)}-\mu_k^{(1)}\right)\right\}\Bigg)\\
&\sim 2c_3c_4\left(2\sum_{k=1}^{N-1} w_{j,k}w_{j,k+1}+\sum_{k=1}^{N-2} w_{j,k}w_{j-1,k+1}+\sum_{k=1}^{N-2} w_{j-1,k}^2\right)\left(\frac{h}{n}\right)f(x)\\
&\quad+\frac{200}{h^2}c_3c_4\left(2\sum_{k=1}^{N-1} w_{j,k}w_{j,k+1}-\sum_{k=1}^{N-2} w_{j,k}w_{j-1,k+1}-\sum_{k=1}^{N-2} w_{j-1,k}^2\right)\left(\frac{h^3}{12n}\right)f(x)\\
&=\left(\frac{245\sqrt{2}-352}{3}\right)c_3c_4\left(\frac{h}{n}\right)f(x).
\end{aligned}
\tag{3.36}
$$

(3.29) $\sim$ (3.36) より、ビン B_j における漸近積分分散 $\mathrm{AIV}\left\{\hat{f}_j(x)\right\}$ は、

$$
\begin{aligned}
\mathrm{AIV}\left\{\hat{f}_j(x)\right\}&=\left\{\frac{6}{5h}+\frac{h^2}{12}\times\frac{120}{7h^3}+\frac{31\sqrt{2}}{6}\times\frac{3}{70h}+2\sqrt{2}\left(2-\sqrt{2}\right)\left(-\frac{1}{10h}\right)\right.\\
&\qquad\left.+\frac{5\sqrt{2}}{3}\left(1-\sqrt{2}\right)h\times\frac{3}{7h^2}+\left(\frac{245\sqrt{2}-352}{3}\right)\frac{1}{280h}\right\}\frac{h}{n}f(x)\\
&=\left(\frac{992+695\sqrt{2}}{840}\right)\frac{1}{n}f(x).
\end{aligned}
\tag{3.37}
$$

したがって、$\mathrm{AIV}\left\{\hat{f}(x)\right\}$ はリーマン積分近似より、

$$\mathrm{AIV}\left\{\hat{f}(x)\right\} = \left(\frac{992+695\sqrt{2}}{840}\right)\frac{1}{nh}. \tag{3.38}$$

3.5 Appendix 2：$\hat{f}_j(x)$ の漸近正規性の証明

節点 $y_k = \frac{k}{N}$, $(k = 1, 2, \ldots, N)$ として、度数 ν_k を定義関数 $I(\cdot)$ を用いて表記すると、

$$\nu_k = \sum_{s=1}^{n} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right), \tag{3.39}$$

ただし、

$$I_s(u,v) = \begin{cases} 1, & u \le X_s \le v \\ 0, & otherwise. \end{cases} \tag{3.40}$$

各ビンの局所 1 次モーメント S_k を定義関数 $I(\cdot)$ を用いて表記すると、

$$S_k = \frac{1}{\nu_k}\sum_{s=1}^{n}(\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right), \tag{3.41}$$

ただし、$I_s(u,v)$ は (3.40) と同様で、x_s, $(s = 1, 2, \ldots, n)$ はデータ点である。

(3.16) を (3.39), (3.41) を用いて書き換えると、

$$\begin{aligned}
\hat{f}_j(x) = {} & \frac{c_1}{n}\sum_{s=1}^{n} I_s\left(\frac{j-1}{N}, \frac{j}{N}\right) \\
& + \frac{c_2}{n}\sum_{s=1}^{n}(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right) \\
& + \frac{c_3}{n}\sum_{k=1}^{N-1} w_{j,k}\left[\sum_{s=1}^{n} I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + \sum_{s=1}^{n} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right. \\
& \qquad \left. - \frac{10}{h}\left\{\sum_{s=1}^{n}(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - \sum_{s=1}^{n}(\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\right] \\
& + \frac{c_4}{n}\sum_{k=1}^{N-1} w_{j-1,k}\left[\sum_{s=1}^{n} I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + \sum_{s=1}^{n} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right. \\
& \qquad \left. - \frac{10}{h}\left\{\sum_{s=1}^{n}(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - \sum_{s=1}^{n}(\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\right],
\end{aligned} \tag{3.42}$$

ただし、$c_1, \ldots, c_4$ について $x - t_j \in \left[-\frac{h}{2}, \frac{h}{2}\right)$ より、有限な値を取る。

(3.42) より、

$$
\begin{aligned}
Z_s = \frac{1}{n}\Bigg(& c_1 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right) \\
& + c_2\left(\mathrm{x}_s - t_j\right) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right) \\
& + c_3 \sum_{k=1}^{N-1} w_{j,k}\Bigg[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right) \\
& - \frac{10}{h}\left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\Bigg] \\
& + c_4 \sum_{k=1}^{N-1} w_{j-1,k}\Bigg[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right) \\
& - \frac{10}{h}\left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\Bigg]\Bigg).
\end{aligned}
\tag{3.43}
$$

このとき、

$$
\hat{f}_j(x) \simeq \sum_{s=1}^{n} Z_s, \tag{3.44}
$$

で、Z_s は独立同一分布に従う。

ここで、

$$
E\left|Z_s - E\left[Z_s\right]\right|^3 \leq E\left|Z_s\right|^3 + 3E\left[Z_s^2\right]\left|E\left[Z_s\right]\right| + 4\left|E\left[Z_s\right]\right|^3. \tag{3.45}
$$

Z_s^2 についてシュワルツの不等式より、

$$
\begin{aligned}
Z_s^2 \leq 6\left(\frac{1}{n}\right)^2 \Bigg[& c_1^2\left\{I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2 \\
& + c_2^2\left\{(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2 \\
& + c_3^2 \sum_{k=1}^{N-1} w_{j,k}^2\left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2 \\
& + \left(\frac{100c_3^2}{h^2}\right) \sum_{k=1}^{N-1} w_{j,k}^2\left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2 \\
& + c_4^2 \sum_{k=1}^{N-1} w_{j-1,k}^2\left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2 \\
& + \left(\frac{100c_4^2}{h^2}\right) \sum_{k=1}^{N-1} w_{j-1,k}^2\left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\Bigg].
\end{aligned}
\tag{3.46}
$$

(3.46) について期待値を取ると、

$$
\begin{aligned}
E\left[Z_s^2\right] \leq 6\left(\frac{1}{n}\right)^2 \Bigg(& c_1^2 E\left[\left\{I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] \\
& + c_2^2 E\left[\left\{(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] \\
& + c_3^2 E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] \\
& + \left(\frac{100c_3^2}{h^2}\right) E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] \\
& + c_4^2 E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2 \left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] \\
& + \left(\frac{100c_4^2}{h^2}\right) E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2 \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right]\Bigg).
\end{aligned}
\tag{3.47}
$$

(3.47) の大括弧内の第 1 項目について、$I_s\left(\frac{k-1}{N}, \frac{k}{N}\right) \overset{i.i.d.}{\sim} B\left(1,\ p_k\right),\ (k = 1, 2, \ldots, N)$ より、

$$
\begin{aligned}
E\left[\left\{I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] &= \mathrm{Var}\left[I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right] + E\left[I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right]^2 \\
&= p_j\left(1 - p_j\right) + p_j^2 \\
&= p_j \\
&= \int_{B_j} f(t)dt \\
&\sim \int_{B_j} \left\{f(x) + (t-x)f'(x) + \cdots\right\} dt \\
&\sim hf(x).
\end{aligned}
\tag{3.48}
$$

したがって、前述の正則条件から $f^{(4)}(x) < \infty$ より、ある有限の値 C_1 を用いて下記の関係が成り立つ;

$$
c_1^2 E\left[\left\{I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] < \frac{C_1}{h}. \tag{3.49}
$$

大括弧内の第 2 項目について、$\mathrm{x}_s - t_k \in \left[-\frac{h}{2},\ \frac{h}{2}\right),\ (k = 1, 2, \ldots, N)$ を用いて、

$$\begin{aligned} E\left[\left\{(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] &= \mathrm{Var}\left[(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right] + E\left[(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right]^2 \\ &\leq h^2\left(\mathrm{Var}\left[I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right] + E\left[I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right]^2\right) \\ &= h^2\left\{p_j\left(1 - p_j\right) + p_j^2\right\} \\ &= h^2 p_j \\ &\sim h^3 f(x). \end{aligned} \tag{3.50}$$

したがって、同様にして、

$$c_2^2 E\left[\left\{(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] < \frac{C_2}{h}. \tag{3.51}$$

大括弧内の第 3 項目について、

$$
\begin{aligned}
&E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] \\
&= \mathrm{Var}\left[\sum_{k=1}^{N-1} w_{j,k} \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\right] \\
&\qquad + E\left[\sum_{k=1}^{N-1} w_{j,k} \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\right]^2 \\
&= \sum_{k=1}^{N-1} w_{j,k}^2 \left(\mathrm{Var}\left[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right)\right] + \mathrm{Var}\left[I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right]\right) \\
&\quad + 2\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j,k} I_s\left(\frac{k}{N}, \frac{k+1}{N}\right), \sum_{k=1}^{N-1} w_{j,k} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right) \\
&\quad + \sum_{k=1}^{N-1} w_{j,k}^2 \left(E\left[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right)\right]^2 + E\left[I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right]^2\right) \\
&\quad + 2\sum_{k=1}^{N-1} w_{j,k}^2 E\left[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right)\right] E\left[I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right] \\
&= \sum_{k=1}^{N-1} w_{j,k}^2 \left\{p_{k+1}\left(1-p_{k+1}\right) + p_k\left(1-p_k\right)\right\} - 2\sum_{k=1}^{N-1} w_{j,k}^2 p_{k+1} p_k \\
&\quad + \sum_{k=1}^{N-1} w_{j,k}^2 \left(p_{k+1}^2 + p_k^2\right) + 2\sum_{k=1}^{N-1} w_{j,k}^2 p_{k+1} p_k \\
&= \sum_{k=1}^{N-1} w_{j,k}^2 \left(p_{k+1} + p_k\right) \\
&\sim 2\sum_{k=1}^{N-1} w_{j,k}^2 hf(x). \qquad (3.52)
\end{aligned}
$$

したがって、同様にして、

$$
c_3^2 E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < \frac{C_3}{h}. \qquad (3.53)
$$

大括弧内の第 4 項目について、

$$
\begin{aligned}
&E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] \\
&\quad = \mathrm{Var}\left[\sum_{k=1}^{N-1} w_{j,k} \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\right] \\
&\qquad\qquad + E\left[\sum_{k=1}^{N-1} w_{j,k} \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\right]^2 \\
&\quad \le h^2 \sum_{k=1}^{N-1} w_{j,k}^2 \left(\mathrm{Var}\left[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right)\right] + \mathrm{Var}\left[I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right]\right) \\
&\quad\quad - 2h^2 \mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j,k} I_s\left(\frac{k}{N}, \frac{k+1}{N}\right), \sum_{k=1}^{N-1} w_{j,k} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right) \\
&\quad\quad + h^2 \sum_{k=1}^{N-1} w_{j,k}^2 \left(E\left[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right)\right] + E\left[I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right]\right)^2 \\
&\quad\quad = \sum_{k=1}^{N-1} w_{j,k}^2 \left(p_{k+1} + p_k\right) \\
&\quad\quad \sim 2 \sum_{k=1}^{N-1} w_{j,k}^2 h f(x). \qquad (3.54)
\end{aligned}
$$

したがって、同様にして、

$$
\left(\frac{100c_3^2}{h^2}\right) E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < \frac{C_4}{h}. \qquad (3.55)
$$

大括弧内第 5, 6 項目は (3.53), (3.55) と同様にして、

$$
c_4^2 E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2 \left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < \frac{C_5}{h}, \qquad (3.56)
$$

$$
\left(\frac{100c_4^2}{h^2}\right) E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2 \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < \frac{C_6}{h}. \qquad (3.57)
$$

以上より、(3.49)〜(3.57) から $E\left[Z_s^2\right]$ は、

$$
E\left[Z_s^2\right] < \frac{6}{n^2} \times \frac{C_1 + \cdots + C_6}{h} \sim O\left(\frac{1}{n^2 h}\right). \qquad (3.58)
$$

また、$|Z_s| = O\left(\frac{1}{nh}\right)$ を用いて、

$$|Z_s|^3 < Z_s^2 O\left(\frac{1}{nh}\right), \tag{3.59}$$

であることから、

$$E|Z_s|^3 = O\left(\frac{1}{n^2h}\right) \times O\left(\frac{1}{nh}\right) = O\left(\frac{1}{n^3h^2}\right). \tag{3.60}$$

したがって、(3.45)、(3.58) より、

$$\begin{aligned}\sum_{s=1}^{n} E|Z_s - E[Z_s]|^3 &= O\left(\frac{1}{n^2h^2}\right) + o\left(\frac{1}{n^2h}\right) + o\left(\frac{1}{n^2}\right) \\ &= O\left(\frac{1}{n^2h^2}\right).\end{aligned} \tag{3.61}$$

$\hat{f}_j(x)$ の分散を $\sigma\left[\hat{f}_j(x)\right]^2$ とすると、$\sigma\left[\hat{f}_j(x)\right]^2 \sim O\left(\frac{1}{nh}\right)$ より、

$$\frac{\sum_{s=1}^{n} E|Z_s - E[Z_s]|^3}{\sigma\left[\hat{f}_j(x)\right]^3} = \frac{O\left(\frac{1}{n^2h^2}\right)}{O\left(\frac{1}{n^{\frac{3}{2}}h^{\frac{3}{2}}}\right)} = O\left(\frac{1}{\sqrt{nh}}\right) = o(1). \tag{3.62}$$

4　平滑化 Polynomial Histogram：S-PH$_{(2,2)}$

本章では、2 次までの局所モーメント情報の利用と節点における累積分布関数の 2 次連続性を同時に満たす多項式 Polynomial Histgram(S-PH) について論じる。

4.1　推定量の構築

3 章と同様に、提案モデルを S-PH$_{(p,q)}$ で表記する。ここで、p は累積分布関数の連続性の次数、q は局所モーメントの次数とする。本章では累積分布関数の 2 次までの連続性条件と、局所 2 次までのモーメント条件を同時に満たす S-PH$_{(2,2)}$ の場合を扱う。

以降、簡単のために記法について説明する。標本数 n、節点数 N、ビン節点 $y_j,\ (j=0,1,\ldots,N)$、定義域 $[y_0,\ y_N]$ とする。j 番目のビン $B_j=[y_{j-1},\ y_j),\ (j=1,2,\ldots,N)$、ビン幅 h、ビンの中点 $t_j=\frac{y_{j-1}+y_j}{2}$、各ビンの面積 $\mu_j^{(0)}$、局所 1 次モーメント $\mu_j^{(1)}$、局所 2 次モーメント $\mu_j^{(2)}$、j 番目のビン B_j における S-PH$_{(2,2)}$ 推定量 $\hat{F}_j(x)$、S-PH$_{(2,2)}$ 密度推定量 $\hat{f}_j(x)=\hat{F}_j'(x)$、$\hat{f}_j(x)$ の係数を $a_0^{(j)},a_1^{(j)},\ldots,a_4^{(j)}$ とし、以降、$a_0,a_1,\ldots,a_4$ と略す。

この時、S-PH$_{(2,2)}$ 密度推定量 $\hat{f}_j(x)$ は次の通りである;

$$\hat{f}_j(x)=a_0Q_0(x)+a_1Q_1(x)+a_2Q_2(x)+a_3Q_3(x)+a_4Q_4(x),\quad x\in B_j, \tag{4.1}$$

ただし、$Q_i(x),\ (i=0,1,\ldots)$ は正規化したルジャンドル多項式で、次の通りである;

$$Q_0(x)=\sqrt{\frac{2}{h}}\sqrt{\frac{1}{2}}, \tag{4.2}$$

$$Q_1(x)=\sqrt{\frac{2}{h}}\frac{\sqrt{6}}{h}\left(x-t_j\right), \tag{4.3}$$

$$Q_2(x)=\sqrt{\frac{2}{h}}\frac{\sqrt{10}}{4}\left\{\frac{12}{h^2}\left(x-t_j\right)^2-1\right\}, \tag{4.4}$$

$$Q_3(x)=\sqrt{\frac{2}{h}}\frac{\sqrt{14}}{2}\left\{\frac{20}{h^3}\left(x-t_j\right)^3-\frac{3}{h}\left(x-t_j\right)\right\}, \tag{4.5}$$

$$Q_4(x)=\sqrt{\frac{2}{h}}\frac{3\sqrt{2}}{16}\left\{\frac{560}{h^4}\left(x-t_j\right)^4-\frac{120}{h^2}\left(x-t_j\right)^2+3\right\}. \tag{4.6}$$

ここで、局所モーメント条件と各ビンの境界での 2 次連続性の条件を満たすために、以下の制約条件を設ける;

(i) 局所 0 次モーメント情報 (面積相等性)：$\int_{B_j}(x-t_j)^0\hat{F}_j'(x)dx=\mu_j^{(0)}$,
(ii) 局所 1 次モーメント情報：$\int_{B_j}(x-t_j)^1\hat{F}_j'(x)dx=\mu_j^{(1)}$,
(iii) 局所 2 次モーメント情報：$\int_{B_j}(x-t_j)^2\hat{F}_j'(x)dx=\mu_j^{(2)}$,
(iv) 累積分布関数の 1 次連続性：$\hat{F}'(y_j-)=\hat{F}'(y_j+)$,
(v) 累積分布関数の 2 次連続性：$\hat{F}''(y_j-)=\hat{F}''(y_j+)$,

ただし、$\hat{F}^{(i)}(y_j-),\ (i=0,1,\dots)$ は節点 y_j における $\hat{F}^{(i)}(x)$ の左方 i 次微分係数、$\hat{F}^{(i)}(y_j+)$ は右方 i 次微分係数である。上記の条件のみでは未知数に対して制約条件の数が 2 つ不足する。そのため、節点 y_j における $\hat{F}_j(x)$ の 2 次微係数 $M_j^{(2)},\ (j=0,1,\dots,N)$ に対し、付加条件として端条件 $M_0^{(2)}=M_N^{(2)}=0$ を仮定する。

各係数について制約条件に基づく方程式を解くと以下を得る;

$$a_0=\frac{\mu_j^{(0)}}{\sqrt{h}}, \tag{4.7}$$

$$a_1=\frac{\sqrt{12h}}{h^2}\mu_j^{(1)}, \tag{4.8}$$

$$a_2=\sqrt{\frac{2}{h}}\left(-\frac{\sqrt{10}}{4}\mu_j^{(0)}+\frac{3\sqrt{10}}{h^2}\mu_j^{(2)}\right), \tag{4.9}$$

$$a_3=\sqrt{\frac{2}{h}}\left\{-\frac{1}{\sqrt{14}h}\mu_j^{(1)}+\frac{\sqrt{14}h^2}{336}\left(M_j^{(2)}+M_{j-1}^{(2)}\right)\right\}, \tag{4.10}$$

$$a_4=\sqrt{\frac{2}{h}}\left\{\frac{\sqrt{2}}{8}\mu_j^{(0)}-\frac{3\sqrt{2}}{2h^2}\mu_j^{(2)}+\frac{\sqrt{2}h^2}{240}\left(M_j^{(2)}-M_{j-1}^{(2)}\right)\right\}, \tag{4.11}$$

ここで、未知の定数 $M_j^{(2)},\ (j=1,2,\dots,N-1)$ について制約条件から行列の形 $\boldsymbol{RM}=\boldsymbol{d}$ で以下の通り表わされる;

$$\begin{pmatrix} 2 & \frac{1}{4} & & & & & \\ \frac{1}{4} & 2 & \frac{1}{4} & & & \mathbf{0} & \\ & \ddots & \ddots & \ddots & & & \\ & & \frac{1}{4} & 2 & \frac{1}{4} & & \\ & & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & & \frac{1}{4} & 2 & \frac{1}{4} \\ & & & & & \frac{1}{4} & 2 \end{pmatrix} \begin{pmatrix} M_1^{(2)} \\ M_2^{(2)} \\ \vdots \\ M_j^{(2)} \\ \vdots \\ M_{N-2}^{(2)} \\ M_{N-1}^{(2)} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_j \\ \vdots \\ d_{N-2} \\ d_{N-1} \end{pmatrix}, \tag{4.12}$$

ただし、$\boldsymbol{M}=\left(M_1^{(2)},\dots,M_{N-1}^{(2)}\right)^\top$, $\boldsymbol{d}$ の要素は $d_j=-\frac{45}{4h^2}\left(\mu_{j+1}^{(0)}-\mu_j^{(0)}\right)-\frac{75}{h^3}\left(\mu_{j+1}^{(1)}+\mu_j^{(1)}\right)+\frac{315}{h^4}\left(\mu_{j+1}^{(2)}-\mu_j^{(2)}\right),\ (j=1,2,\dots,N-1)$ である。節点数 N が十分大きく、$M_j^{(2)}$ は定義域の端点から十分に離れたビンでの要素とすると、

$$M_j^{(2)}=\frac{3}{h^2}\sum_{k=1}^{N-1}w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)}-\mu_k^{(0)}\right)-\frac{10}{h}\left(\mu_{k+1}^{(1)}+\mu_k^{(1)}\right)+\frac{42}{h^2}\left(\mu_{k+1}^{(2)}-\mu_k^{(2)}\right)\right\}, \tag{4.13}$$

ただし、$w_{j,k}=\frac{\sqrt{15}}{3}\left(\sqrt{15}-4\right)^{|j-k|}$、$\sum_{k=1}^{N-1}\frac{\sqrt{15}}{3}\left(\sqrt{15}-4\right)^{|j-k|}=1$ である。

以上より、S-PH$_{(2,2)}$ の密度推定量 $\hat{f}_j(x)$ は次式で与えられる;

$$\begin{aligned}\hat{f}_j(x) = \sqrt{\frac{2}{h}}\Bigg[&\left(\frac{1}{\sqrt{2}}Q_0(x) - \frac{\sqrt{10}}{4}Q_2(x) + \frac{\sqrt{2}}{8}Q_4(x)\right)\mu_j^{(0)} \\ &+ \left(\frac{\sqrt{6}}{h}Q_1(x) - \frac{1}{\sqrt{14}h}Q_3(x)\right)\mu_j^{(1)} + \left(\frac{3\sqrt{10}}{h^2}Q_2(x) - \frac{3\sqrt{2}}{2h^2}Q_4(x)\right)\mu_j^{(2)} \\ &+ \left(\frac{\sqrt{14}}{112}Q_3(x) + \frac{\sqrt{2}}{80}Q_4(x)\right) \\ &\qquad \times \sum_{k=1}^{N-1} w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\} \\ &+ \left(\frac{\sqrt{14}}{112}Q_3(x) - \frac{\sqrt{2}}{80}Q_4(x)\right) \\ &\qquad \times \sum_{k=1}^{N-1} w_{j-1,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\}\Bigg].\end{aligned} \tag{4.14}$$

4.2 漸近的性質

本節ではビンごとに推定された $\hat{f}_j(x)$ で構成された S-PH$_{(2,2)}$ 密度推定量の漸近的 MISE を導き、漸近正規性が成り立つことを示す。

S-PH$_{(2,2)}$ 密度推定量に関して次の正則条件を満たすものとする;

(i) ビン幅 h について、$n \to \infty$ のとき、$h \to 0$ かつ $nh \to \infty$,
(ii) 関数 $f^{(5)}(x)$ は絶対連続関数で、$R\left(f^{(6)}\right) = \int f^{(6)}(x)^2 dx < \infty$.

この (i) と (ii) の条件のもとで、以下の S-PH$_{(2,2)}$ 推定量の漸近分散、漸近バイアス、漸近的な MISE と漸近正規性を以下の定理及び系にまとめる。

Theorem 1 : S-PH$_{(2,2)}$ の IV$\left\{\hat{f}_j(x)\right\}$

ビン B_j における S-PH$_{(2,2)}$ 密度推定量の IV$\left\{\hat{f}_j(x)\right\}$ は、

$$\mathrm{IV}\left\{\hat{f}_j(x)\right\} = \sum_{i=0}^{4} \mathrm{Var}(c_i), \tag{4.15}$$

ただし、c_i, $(i = 0, 1, \ldots, 4)$ は (4.14) 式について $Q_i(x)$ で整理した時の係数であり、詳しくは 4.4 節の Appendix 1 の (4.26)〜(4.30) に対応している。

Theorem 2：S-PH$_{(2,2)}$ の Bias$\left\{\hat{f}_j(x)\right\}$ 及び ISB$\left\{\hat{f}_j(x)\right\}$

ビン B_j における S-PH$_{(2,2)}$ 密度推定量の Bias$\left\{\hat{f}_j(x)\right\}$ 及び ISB$\left\{\hat{f}_j(x)\right\}$ は、

$$\mathrm{Bias}\left\{\hat{f}_j(x)\right\} = h^6 f^{(5)}(x)\left(-\frac{1}{332640}\sqrt{\frac{11}{h}}Q_5(x) + \frac{1}{211680}\sqrt{\frac{7}{h}}Q_3(x)\right), \tag{4.16}$$

$$\mathrm{ISB}\left\{\hat{f}_j(x)\right\} = \left\{\left(-\frac{\sqrt{11}}{332640}\right)^2 + \left(\frac{\sqrt{7}}{211680}\right)^2\right\} h^{11} f^{(5)}(x)^2. \tag{4.17}$$

Theorem 3：S-PH$_{(2,2)}$ の AMISE

S-PH$_{(2,2)}$ 密度推定量 $\hat{f}_{SPH(2,2)}(x)$ の漸近的な MISE(AMISE) は、

$$\begin{aligned}\mathrm{AMISE}[\hat{f}_{SPH(2,2)}(x)] &= \left(\frac{13425+2558\sqrt{15}}{6300}\right)\frac{1}{nh} + \frac{h^{10}}{3911846400}R\left(f^{(5)}\right) \\ &\fallingdotseq \frac{3.703}{nh} + \frac{h^{10}}{3911846400}R\left(f^{(5)}\right).\end{aligned} \tag{4.18}$$

このときの最適ビン幅 $h^*_{SPH(2,2)}$ 及び最小 AMISE$_{SPH(2,2)}$ は、

$$h^*_{SPH(2,2)} = \left\{\frac{310464(13425+2558\sqrt{15})}{5}\right\}^{\frac{1}{11}} R\left(f^{(5)}\right)^{-\frac{1}{11}} n^{-\frac{1}{11}}, \tag{4.19}$$

$$\text{最小 AMISE}_{SPH(2,2)} = \frac{11}{10}\left(\frac{13425+2558\sqrt{15}}{6300}\right)^{\frac{10}{11}}\left(\frac{R\left(f^{(5)}\right)}{391184640}\right)^{\frac{1}{11}} n^{-\frac{10}{11}}. \tag{4.20}$$

Theorem 4：S-PH$_{(2,2)}$ の漸近正規性

S-PH$_{(2,2)}$ 密度推定量について、

$$\frac{\sum_{s=1}^{n} E\left|Z_s - E\left[Z_s\right]\right|^3}{\sigma\left[\hat{f}_j(x)\right]^3} = \frac{O\left(\frac{1}{n^2h^2}\right)}{O\left(\frac{1}{n^{3/2}h^{3/2}}\right)} = O\left(\frac{1}{n^{1/2}h^{1/2}}\right) = o(1), \tag{4.21}$$

ただし、$\sigma\left[\hat{f}_j(x)\right]^2 = \mathrm{Var}\left\{\hat{f}_j(x)\right\}$ で、$Z_s,\ (s=1,2,\ldots,n)$ は $\hat{f}_j(x)$ に従う独立な確率変数である。したがって、リアプノフの条件を満たすことから漸近正規性が成り立つ。

Corollary 1：各ビンにおける S-PH$_{(2,2)}$ の漸近正規性

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

$\alpha = \frac{1}{11}$ のとき、

$$\sqrt{nh}\left\{\hat{f}_j(x) - f(x)\right\} \xrightarrow{d} N\left(\mathrm{Bias}\left[\hat{f}_j(x)\right], \left(\frac{13425+2558\sqrt{15}}{6300}\right) f\left(\xi_j\right)\right), \tag{4.22}$$

$\alpha > \frac{1}{11}$ のとき、

$$\sqrt{nh}\left\{\hat{f}_j(x) - f(x)\right\} \xrightarrow{d} N\left(o(1),\ \left(\frac{13425 + 2558\sqrt{15}}{6300}\right) f\left(\xi_j\right)\right), \tag{4.23}$$

が漸近的に成り立つ。ただし、$f(\xi_j)$ は $p_j = \int_{B_j} f(t)dt = hf(\xi_j),\ \xi_j \in B_j$ を満たす B_j 内のある点とする。Theorem1〜3 の S-PH$_{(2,2)}$ の IV、Bias、ISB、AMISE については 4.4 節の Appendix 1、Theorem 4 の漸近正規性の証明については 4.5 節の Appendix 2 で詳述する。

4.3 数値実験

S-PH$_{(2,2)}$ の有限標本における推定精度を調べるため、積分二乗誤差 (ISE) について数値実験を行う。定義域 $[-7,\ 7]$ の混合正規分布 $f(x)$ は $\frac{4}{5}N\left(-2, 2^2\right) + \frac{1}{5}N\left(4, 1.5^2\right)$ に従う確率密度関数とし、データ数を $n = 10^2,\ 10^3,\ 10^4,\ 10^5$ とする。ビン幅は理論上の最適ビン幅を用いて、各データ数における ISE の数値実験 1000 回の平均 (MISE) と標準偏差を算出する。同様の設定の下での Histogram、Histospline、2 次 PH、S-PH$_{(2,1)}$ と比較する。

数値実験に用いる各推定量の最適ビン幅 h^* は次の通りである。

表 4.1　各推定量の最適ビン幅

Histogram	$h^*_{HIST} = \left(\frac{6}{R(f^{(1)})}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}$
Histospline	$h^*_{HSP} = \left(\frac{2520\sqrt{3}+1512}{R(f^{(3)})}\right)^{\frac{1}{7}} n^{-\frac{1}{7}}$
2 次 PH	$h^*_{PH(2)} = \left(\frac{50400}{R(f^{(3)})}\right)^{\frac{1}{7}} n^{-\frac{1}{7}}$
S-PH$_{(2,1)}$	$h^*_{SPH(2,1)} = \left(\frac{125100\sqrt{2}+178560}{R(f^{(4)})}\right)^{\frac{1}{9}} n^{-\frac{1}{9}}$
S-PH$_{(2,2)}$	$h^*_{SPH(2,2)} = \left\{\frac{310464(13425+2558\sqrt{15})}{5R\left(f^{(5)}\right)}\right\}^{\frac{1}{11}} n^{-\frac{1}{11}}$

図 4.1 は $n = 10^3$ の時の Histogram、Histospline、2 次 PH と提案モデル S-PH$_{(2,2)}$ の数値実験結果の例で、実線が推定結果、破線が真の密度関数のグラフである。

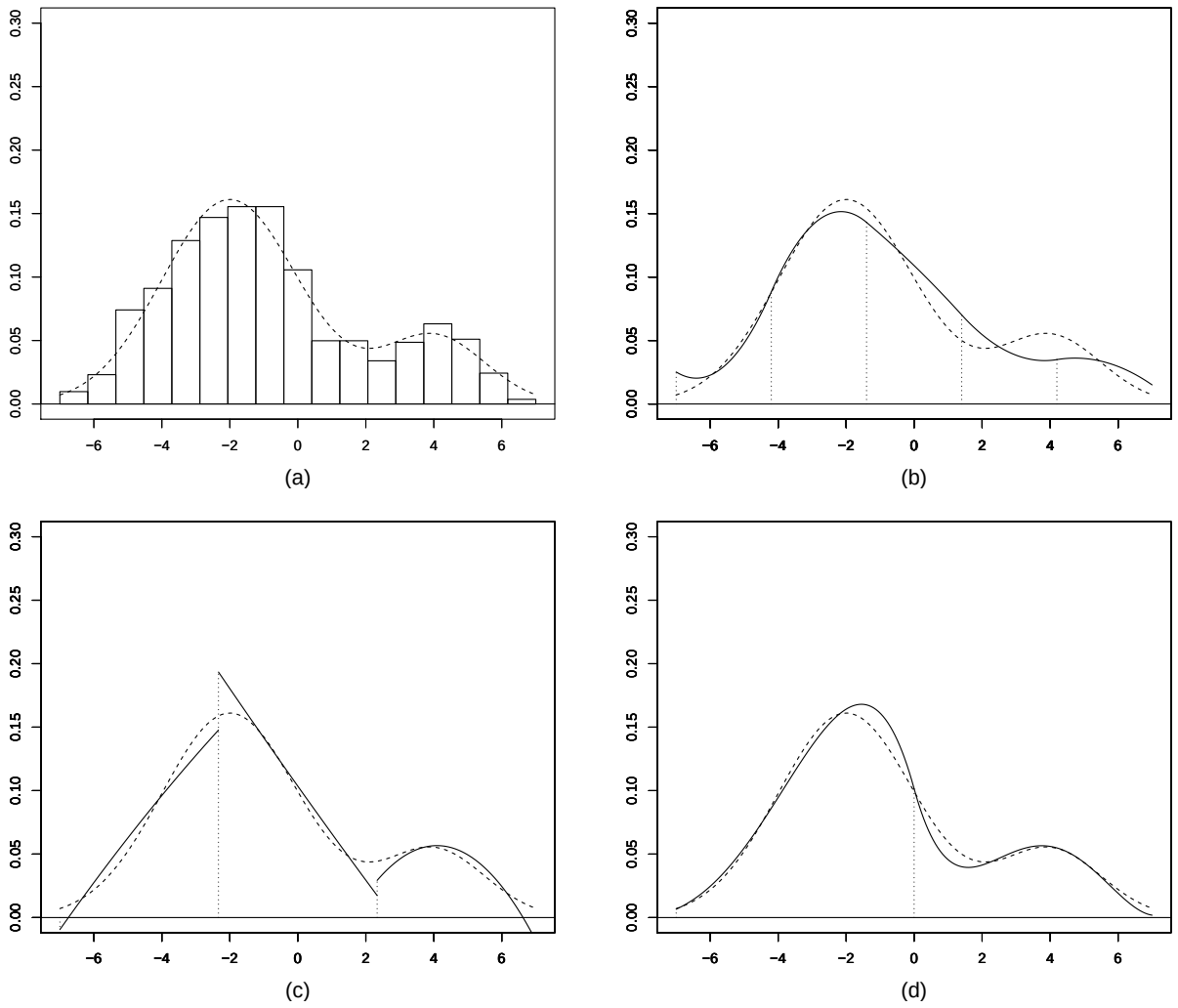

図 4.1 $n = 10^3$ での推定例 (実線が推定結果、破線が真の分布。(a)Histogram、(b)Histospline、(c)2 次 PH、(d)S-PH$_{(2,2)}$)

表 4.2 主な 4 つの推定量と S-PH$_{(2,2)}$ の MISE 比較 (繰り返し回数 1000 回)

	MISE*	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
Histogram	$O\left(n^{-\frac{2}{3}}\right)$	0.00745	0.00172	0.00039	0.00009
Histospline	$O\left(n^{-\frac{6}{7}}\right)$	0.01850	0.00154	0.00011	0.00003
2 次 PH	$O\left(n^{-\frac{6}{7}}\right)$	<u>0.00554</u>	0.00092	0.00012	<u>0.00002</u>
S-PH$_{(2,1)}$	$O\left(n^{-\frac{8}{9}}\right)$	0.00570	<u>0.00057</u>	0.00017	<u>0.00002</u>
S-PH$_{(2,2)}$	$O\left(n^{-\frac{10}{11}}\right)$	0.01396	0.00060	<u>0.00008</u>	<u>0.00002</u>

表 4.2 はデータ数 n を変化させた時に、5 つの推定法を比較した MISE の実験結果である。n ごとに比較して最小の値に下線を引いてある。どの推定量も標本数が大きくなるにつれて ISE は小さくなる。標本数が $n = 10^4$ 以上の時に、S-PH$_{(2,2)}$ の ISE が最も小さい。$n = 10^5$ の時、Histogram 以外の推定法の優劣はほとんどない。

表 4.3　データ数 n と最適ビン数の比較

	MISE*	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
Histogram	$O\left(n^{-\frac{2}{3}}\right)$	7	17	36	78
Histospline	$O\left(n^{-\frac{6}{7}}\right)$	3	5	7	9
2 次 PH	$O\left(n^{-\frac{6}{7}}\right)$	2	3	5	7
S-PH$_{(2,1)}$	$O\left(n^{-\frac{8}{9}}\right)$	3	4	5	6
S-PH$_{(2,2)}$	$O\left(n^{-\frac{10}{11}}\right)$	<u>1</u>	<u>2</u>	<u>3</u>	<u>3</u>
5 次 Kernel	$O\left(n^{-\frac{10}{11}}\right)$	10^2	10^3	10^4	10^5

表 4.3 は標本数 $n = 10^{2\sim5}$ と変化させたときの Histogram、2 次 PH、Histospline、S-PH$_{(2,1)}$、S-PH$_{(2,2)}$、5 次 Kernel 推定のビン数である。ただし、5 次 Kernel は各データ点で推定するため、データ数をビン数に対応させている。n を変化させた時の最小の値に下線を引いてある。標本数に関わらず S-PH$_{(2,2)}$ のビン数が他の推定法と比較して最も小さい。例えば、$n = 10^5$ では Histogram の約 1/25、5 次 Kernel 推定の約 1/33000 のビン数である。すなわち、S-PH$_{(2,2)}$ は他の推定法と比較して最も少ないビン数で推定できる。

表 4.4　主な 4 つの推定量と S-PH$_{(2,2)}$ の ISE 標準偏差比較 (繰り返し回数 1000 回)

	MISE*	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$
Histogram	$O\left(n^{-\frac{2}{3}}\right)$	0.002678	0.001718	0.000072	0.000011
Histospline	$O\left(n^{-\frac{6}{7}}\right)$	0.002058	0.000292	<u>0.000036</u>	<u>0.000005</u>
2 次 PH	$O\left(n^{-\frac{6}{7}}\right)$	0.002238	0.000423	0.000043	0.000006
S-PH$_{(2,1)}$	$O\left(n^{-\frac{8}{9}}\right)$	0.002982	0.000333	0.000043	<u>0.000005</u>
S-PH$_{(2,2)}$	$O\left(n^{-\frac{10}{11}}\right)$	<u>0.001120</u>	<u>0.000288</u>	0.000042	<u>0.000005</u>

表 4.4 はデータ数 n を変化させた時の 5 つの推定法の安定性について、ISE の標準偏差を比較した実験結果である。n ごとに比較して最小の値に下線を引いてある。どの推定量も標本数が大きく

なるにつれて ISE の標準偏差は小さくなる。$n = 10^4$ 以外では、S-PH$_{(2,2)}$ は ISE の標準偏差が最も小さく、安定した推定法である。

以上から、表 4.2 と表 4.4 より n が大きくなるにつれて誤差が小さくなる一致性の裏付けとともに、本稿で提案した S-PH$_{(2,2)}$ は総じて理論的に優れていることが数値実験においても裏付けられ、ビン幅が大きい特性から安定性が得られている。

4.4 Appendix 1：AMISE$\{\hat{f}(x)\}$ の証明

S-PH$_{(2,2)}$ 推定量の AMISE について漸近積分分散 AIV$\left\{\hat{f}(x)\right\}$ と漸近積分二乗バイアス AISB$\left\{\hat{f}(x)\right\}$ のそれぞれから導出する。

まずビン B_j における AIV$\left\{\hat{f}_j(x)\right\}$ について示す。S-PH$_{(2,2)}$ の密度推定量は (4.14) 式より、

$$
\begin{aligned}
\hat{f}_j(x) = \sqrt{\frac{2}{h}}\Bigg(&\frac{\mu_j^{(0)}}{\sqrt{2}}Q_0(x) + \frac{\sqrt{6}}{h}\mu_j^{(1)}Q_1(x) + \left(-\frac{\sqrt{10}}{4}\mu_j^{(0)} + \frac{3\sqrt{10}}{h^2}\mu_j^{(2)}\right)Q_2(x) \\
&+ \Bigg[-\frac{\mu_j^{(1)}}{\sqrt{14}h} + \frac{\sqrt{14}}{112}\sum_{k=1}^{N-1} w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\} \\
&\quad + \frac{\sqrt{14}}{112}\sum_{k=1}^{N-1} w_{j-1,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\}\Bigg]Q_3(x) \\
&+ \Bigg[\frac{\sqrt{2}}{8}\mu_j^{(0)} - \frac{3\sqrt{2}}{2h^2}\mu_j^{(2)} \\
&\quad + \frac{\sqrt{2}}{80}\sum_{k=1}^{N-1} w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\} \\
&\quad - \frac{\sqrt{2}}{80}\sum_{k=1}^{N-1} w_{j-1,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\}\Bigg]Q_4(x)\Bigg).
\end{aligned}
\tag{4.24}
$$

(4.24) より、IV$\left\{\hat{f}_j(x)\right\}$ は以下の通り表現される;

$$
\mathrm{IV}\left\{\hat{f}_j(x)\right\} = \sum_{i=0}^{4}\mathrm{Var}(c_i)\int_{B_j} Q_i(x)^2dx + \sum_{i\neq l}^{4}\mathrm{Cov}\left(c_i, c_l\right)\int_{B_j} Q_i(x)Q_l(x)dx, \tag{4.25}
$$

ただし、

$$c_0 = \frac{\mu_j^{(0)}}{\sqrt{h}}, \tag{4.26}$$

$$c_1 = \frac{2\sqrt{3}}{h^{3/2}}\mu_j^{(1)}, \tag{4.27}$$

$$c_2 = \sqrt{\frac{2}{h}}\left(-\frac{\sqrt{10}}{4}\mu_j^{(0)} + \frac{3\sqrt{10}}{h^2}\mu_j^{(2)}\right), \tag{4.28}$$

$$\begin{aligned} c_3 = \sqrt{\frac{2}{h}}\Bigg[& -\frac{\mu_j^{(1)}}{\sqrt{14}h} \\ & + \frac{\sqrt{14}}{112}\sum_{k=1}^{N-1} w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\} \\ & + \frac{\sqrt{14}}{112}\sum_{k=1}^{N-1} w_{j-1,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\}\Bigg], \end{aligned} \tag{4.29}$$

$$\begin{aligned} c_4 = \sqrt{\frac{2}{h}}\Bigg[& -\frac{\mu_j^{(1)}}{\sqrt{14}h} \\ & + \frac{\sqrt{2}}{80}\sum_{k=1}^{N-1} w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\} \\ & + \frac{\sqrt{2}}{80}\sum_{k=1}^{N-1} w_{j-1,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\}\Bigg]. \end{aligned} \tag{4.30}$$

(4.25) 式の第 1 項について、正規性から $\int_{B_j} Q_i(x)^2 dx = 1$ となる。第 2 項について、共分散行列の非対角成分要素が直交性から 0 となる。そのため、(4.25) 式の第 1 項のみ評価すればよい。以降での分散の評価では、主要項のみ取り出して記述する。

(4.25) の第 1 項は、各ビンの度数を ν_k, $(k = 1, 2, \ldots, N)$ とし、$\mu_j^{(0)} = \frac{\nu_j}{n}$ を用いると、近似的に、

$$\begin{aligned} \mathrm{Var}\left(c_0 Q_0(x)\right) &= \mathrm{Var}\left(\frac{\mu_j^{(0)}}{\sqrt{h}}Q_0(x)\right) \\ &= \frac{Q_0(x)^2}{n^2 h}\mathrm{Var}\left(\nu_j^{(0)}\right) \\ &= \frac{Q_0(x)^2}{n^2 h} n p_j (1 - p_j) \\ &\sim \frac{1}{n} f(x) Q_0(x)^2. \end{aligned} \tag{4.31}$$

(4.25) の第 2 項は、同様にして、

$$\begin{aligned}\mathrm{Var}\left(c_1 Q_1(x)\right) &= \mathrm{Var}\left(\frac{2\sqrt{3}}{h^{3/2}}\mu_j^{(1)} Q_1(x)\right) \\ &= \frac{12}{h^3} Q_1(x)^2 \mathrm{Var}\left(\mu_j^{(1)}\right) \\ &\sim \frac{1}{n} f(x) Q_1(x)^2. \end{aligned} \tag{4.32}$$

(4.25) の第 3 項は、同様にして、

$$\begin{aligned}\mathrm{Var}\left(c_2 Q_2(x)\right) &= \mathrm{Var}\left\{\sqrt{\frac{2}{h}}\left(-\frac{\sqrt{10}}{4}\mu_j^{(0)} + \frac{3\sqrt{10}}{h^2}\mu_j^{(2)}\right) Q_2(x)\right\} \\ &= \frac{2}{h} Q_2(x)^2 \left\{\frac{5}{8n^2}\mathrm{Var}\left(\nu_j\right) + \frac{90}{h^2}\mathrm{Var}\left(\mu_j^{(2)}\right) - \frac{15}{h^2}\mathrm{Cov}\left(\mu_j^{(0)}, \mu_j^{(2)}\right)\right\} \\ &\sim \frac{2}{h} Q_2(x)^2 \left(\frac{5h}{8n} f(x) + \frac{9h}{8n} f(x) - \frac{5h}{4n} f(x)\right) \\ &= \frac{1}{n} f(x) Q_2(x)^2. \end{aligned} \tag{4.33}$$

(4.25) の第 4 項は、$\sum_{k=1}^{N-2} w_{j,k} w_{j,k+1} = \sum_{k=1}^{N-2} w_{j-1,k} w_{j-1,k+1} = -\frac{\sqrt{15}}{9}$,
$\sum_{k=1}^{N-1} w_{j,k}^2 = \sum_{k=1}^{N-1} w_{j-1,k}^2 = \frac{4\sqrt{15}}{9}$, $\sum_{k=1}^{N-2} w_{j,k} w_{j-1,k+1} = \frac{450-116\sqrt{15}}{9}$ を用いて、同様にして、

$$
\begin{aligned}
&\mathrm{Var}\left(c_3 Q_3(x)\right) \\
&= \mathrm{Var}\Bigg(\sqrt{\frac{2}{h}}\Bigg[-\frac{\mu_j^{(1)}}{\sqrt{14}h} + \frac{\sqrt{14}}{112}\sum_{k=1}^{N-1} w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\} \\
&\qquad + \frac{\sqrt{14}}{112}\sum_{k=1}^{N-1} w_{j-1,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\}\Bigg] Q_3(x)\Bigg) \\
&\sim \frac{2}{h} Q_3(x)^2 \Bigg\{ \frac{1}{14h^2}\mathrm{Var}\left(\mu_j^{(1)}\right) + \frac{9}{3584}\mathrm{Var}\left(\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right)\right) \\
&\qquad + \frac{25}{224h^2}\mathrm{Var}\left(\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right)\right) + \frac{63}{32h^4}\mathrm{Var}\left(\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right) \\
&\qquad + \frac{9}{3584}\mathrm{Var}\left(\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right)\right) + \frac{25}{224h^2}\mathrm{Var}\left(\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right)\right) \\
&\qquad + \frac{63}{32h^4}\mathrm{Var}\left(\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right) \\
&\qquad - \frac{9}{64h^2}\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right), \sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right) \\
&\qquad - \frac{9}{64h^2}\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right), \sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right) \\
&\qquad + \frac{9}{1792}\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right), \sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right)\right) \\
&\qquad + \frac{25}{112h^2}\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right), \sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right)\right) \\
&\qquad + \frac{63}{16h^4}\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right), \sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right) \\
&\qquad - \frac{9}{64h^2}\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right), \sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right) \\
&\qquad - \frac{9}{64h^2}\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right), \sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right) \\
&\qquad + \frac{5}{28h^2}\mathrm{Cov}\left(\mu_j^{(1)}, \sum_{k=1}^{N-1} w_{j,k}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right)\right) + \frac{5}{28h^2}\mathrm{Cov}\left(\mu_j^{(1)}, \sum_{k=1}^{N-1} w_{j-1,k}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right)\right)\Bigg\} \\
&\sim \left(\frac{113\sqrt{15} - 306}{336}\right)\frac{1}{n} f(x) Q_3(x)^2. \qquad (4.34)
\end{aligned}
$$

(4.25) の第 5 項は、 (4.34) 式と同様にして、

$$
\begin{aligned}
&\mathrm{Var}\left(c_4 Q_4(x)\right) \\
&= \mathrm{Var}\Bigg(\sqrt{\frac{2}{h}}\Bigg[-\frac{\mu_j^{(1)}}{\sqrt{14}h} + \frac{\sqrt{2}}{80}\sum_{k=1}^{N-1} w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\} \\
&\qquad + \frac{\sqrt{2}}{80}\sum_{k=1}^{N-1} w_{j-1,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\}\Bigg] Q_4(x)\Bigg) \\
&\sim \left(\frac{150 + 251\sqrt{15}}{3600}\right)\frac{1}{n} f(x) Q_4(x)^2. \qquad (4.35)
\end{aligned}
$$

以上の (4.31) ∼ (4.35) より、ビン B_j における漸近積分分散 $\mathrm{AIV}\left\{\hat{f}_j(x)\right\}$ は、

$$
\begin{aligned}
\mathrm{AIV}\left\{\hat{f}_j(x)\right\} &= \Bigg(\int_{B_j} Q_0(x)^2 dx + \int_{B_j} Q_1(x)^2 dx + \int_{B_j} Q_2(x)^2 dx \\
&\qquad + \frac{113\sqrt{15} - 306}{336}\int_{B_j} Q_3(x)^2 dx + \frac{251\sqrt{15} + 150}{3600}\int_{B_j} Q_4(x)^2 dx\Bigg)\frac{1}{n} f(x) \\
&= \left(\frac{2558\sqrt{15} + 13425}{6300}\right)\frac{1}{n} f(x). \qquad (4.36)
\end{aligned}
$$

したがって、$\mathrm{AIV}\left\{\hat{f}(x)\right\}$ はリーマン積分近似 $\sum_j f(\xi_j)h = \left[\int f(x)dx + o(1)\right]$ より、

$$
\mathrm{AIV}\left\{\hat{f}(x)\right\} = \left(\frac{2558\sqrt{15} + 13425}{6300}\right)\frac{1}{nh}. \qquad (4.37)
$$

次に、漸近積分二乗バイアスについて示す。(4.24) を整理し直すと、

$$
\begin{aligned}
\hat{f}_j(x) = \sqrt{\frac{2}{h}}\Bigg[&\left(\frac{1}{\sqrt{2}}Q_0(x) - \frac{\sqrt{10}}{4}Q_2(x) + \frac{\sqrt{2}}{8}Q_4(x)\right)\mu_j^{(0)} \\
&+ \left(\frac{\sqrt{6}}{h}Q_1(x) - \frac{1}{\sqrt{14}h}Q_3(x)\right)\mu_j^{(1)} + \left(\frac{3\sqrt{10}}{h^2}Q_2(x) - \frac{3\sqrt{2}}{2h^2}Q_4(x)\right)\mu_j^{(2)} \\
&+ \left(\frac{\sqrt{14}}{112}Q_3(x) + \frac{\sqrt{2}}{80}Q_4(x)\right) \\
&\qquad \times \sum_{k=1}^{N-1} w_{j,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\} \\
&+ \left(\frac{\sqrt{14}}{112}Q_3(x) - \frac{\sqrt{2}}{80}Q_4(x)\right) \\
&\qquad \times \sum_{k=1}^{N-1} w_{j-1,k}\left\{-\frac{3}{2}\left(\mu_{k+1}^{(0)} - \mu_k^{(0)}\right) - \frac{10}{h}\left(\mu_{k+1}^{(1)} + \mu_k^{(1)}\right) + \frac{42}{h^2}\left(\mu_{k+1}^{(2)} - \mu_k^{(2)}\right)\right\}. \qquad (4.38)
\end{aligned}
$$

(4.38) 式の期待値を取ると、

$$E\left[\hat{f}_j(x)\right] = \sqrt{\frac{2}{h}}\Bigg[\gamma_0 E\left[\mu_j^{(0)}\right] + \gamma_1 E\left[\mu_j^{(1)}\right] + \gamma_2 E\left[\mu_j^{(2)}\right] + \gamma_3 \sum_{k=1}^{N-1} w_{j,k}\Bigg\{-\frac{3}{2}\left(E\left[\mu_{k+1}^{(0)}\right] - E\left[\mu_k^{(0)}\right]\right) - \frac{10}{h}\left(E\left[\mu_{k+1}^{(1)}\right] + E\left[\mu_k^{(1)}\right]\right) + \frac{42}{h^2}\left(E\left[\mu_{k+1}^{(2)}\right] - E\left[\mu_k^{(2)}\right]\right)\Bigg\} + \gamma_4 \sum_{k=1}^{N-1} w_{j-1,k}\Bigg\{-\frac{3}{2}\left(E\left[\mu_{k+1}^{(0)}\right] - E\left[\mu_k^{(0)}\right]\right) - \frac{10}{h}\left(E\left[\mu_{k+1}^{(1)}\right] + E\left[\mu_k^{(1)}\right]\right) + \frac{42}{h^2}\left(E\left[\mu_{k+1}^{(2)}\right] - E\left[\mu_k^{(2)}\right]\right)\Bigg\}\Bigg], \tag{4.39}$$

ただし、

$$\gamma_0 = \frac{1}{\sqrt{2}}Q_0(x) - \frac{\sqrt{10}}{4}Q_2(x) + \frac{\sqrt{2}}{8}Q_4(x), \tag{4.40}$$

$$\gamma_1 = \frac{\sqrt{6}}{h}Q_1(x) - \frac{1}{\sqrt{14}h}, \tag{4.41}$$

$$\gamma_2 = \frac{3\sqrt{10}}{h^2}Q_2(x) - \frac{3\sqrt{2}}{2h^2}Q_4(x), \tag{4.42}$$

$$\gamma_3 = \frac{\sqrt{14}}{112}Q_3(x) + \frac{\sqrt{2}}{80}Q_4(x), \tag{4.43}$$

$$\gamma_4 = \frac{\sqrt{14}}{112}Q_3(x) - \frac{\sqrt{2}}{80}Q_4(x). \tag{4.44}$$

$\nu_k \sim B\left(n,\ p_k\right),\ (k = 1, 2, \ldots, N)$ で、$p_k = \int_{B_k} f(t)dt$ の時、

$$E\left[\mu_k^{(0)}\right] = \frac{1}{n}E\left[\nu_k\right] = \frac{1}{n}np_k = \int_{B_k} f(t)dt, \tag{4.45}$$

$$E\left[\mu_k^{(1)}\right] = \int_{B_k} (t - t_k)f(t)dt, \tag{4.46}$$

$$E\left[\mu_k^{(2)}\right] = \int_{B_k} (t - t_k)^2 f(t)dt, \tag{4.47}$$

ここで、未知の $f(t)$ についてテイラー級数による近似から以下の通り表せられる;

$$\int_{B_k} f(t)dt \sim \int_{B_k} \Bigg\{f(x) + (t-x)f^{(1)}(x) + \frac{1}{2!}(t-x)^2 f^{(2)}(x) + \frac{1}{3!}(t-x)^3 f^{(3)}(x) + \frac{1}{4!}(t-x)^4 f^{(4)}(x) + \frac{1}{5!}(t-x)^5 f^{(5)}(x) + \cdots\Bigg\}dt. \tag{4.48}$$

(4.39) 式で (4.40)〜(4.48) を用いて、$\sum_{k=1}^{N-1} w_{j,k}k = j$, $\sum_{k=1}^{N-1} w_{j,k}k^2 = j^2 - \frac{1}{5}$, $\sum_{k=1}^{N-1} w_{j,k}k^3 = j^3 - \frac{3}{5}j$, $\sum_{k=1}^{N-1} w_{j,k}k^4 = j^4 - \frac{6}{5}j^2 + \frac{1}{25}$ であることを利用して整理す

ると、$\hat{f}_j(x)$ のバイアスは、

$$\begin{aligned}\mathrm{Bias}\left[\hat{f}_j(x)\right] &= E\left[\hat{f}_j(x)\right] - f(x) \\ &= h^6 f^{(5)}(x)\left(-\frac{1}{332640}\sqrt{\frac{11}{h}}Q_5(x) + \frac{1}{211680}\sqrt{\frac{7}{h}}Q_3(x)\right). \end{aligned} \tag{4.49}$$

したがって、ビン B_j における $\mathrm{AISB}\left\{\hat{f}_j(x)\right\}$ は、

$$\begin{aligned}\mathrm{AISB}\left\{\hat{f}_j(x)\right\} &= h^{12} f^{(5)}(x)^2 \left\{\left(\frac{1}{332640}\right)^2 \frac{11}{h}\int_{B_j} Q_5(x)^2 dx + \left(\frac{1}{211680}\right)^2 \frac{7}{h}\int_{B_j} Q_3(x)^2 dx \right. \\ &\qquad \left. - 2\left(\frac{1}{332640}\sqrt{\frac{11}{h}}\right)\left(\frac{1}{211680}\sqrt{\frac{7}{h}}\right)\int_{B_j} Q_5(x)Q_3(x)dx\right\} \\ &= \frac{h^{11}}{3911846400} f^{(5)}(x)^2. \end{aligned} \tag{4.50}$$

以上より、$\mathrm{AISB}\left\{\hat{f}(x)\right\}$ はリーマン積分近似 $\sum_j f^{(5)}(\xi_j)^2 h = \left[\int f^{(5)}(x)^2 dx + o(1)\right]$ を用いて、

$$\mathrm{AISB}\left\{\hat{f}(x)\right\} = \frac{R\left(f^{(5)}\right)}{3911846400} h^{10}. \tag{4.51}$$

4.5 Appendix 2：$\hat{f}_j(x)$ の漸近正規性の証明

節点 $y_k = \frac{k}{N}$, $(k = 1, 2, \ldots, N)$ として、度数 ν_k を定義関数 $I(\cdot)$ を用いて表記すると、

$$\nu_k = \sum_{s=1}^{n} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right), \tag{4.52}$$

ただし、

$$I_s(u, v) = \begin{cases} 1, & u \le X_s \le v \\ 0, & otherwise. \end{cases} \tag{4.53}$$

各ビンの局所 1 次モーメント $S_k^{(1)}$、局所 2 次モーメント $S_k^{(2)}$ を定義関数 $I(\cdot)$ を用いて表記すると、

$$S_k^{(1)} = \frac{1}{\nu_k}\sum_{s=1}^{n}(\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right), \tag{4.54}$$

$$S_k^{(2)} = \frac{1}{\nu_k}\sum_{s=1}^{n}(\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right), \tag{4.55}$$

ただし、$I_s(u, v)$ は (4.53) と同様で、x_s, $(s = 1, 2, \ldots, n)$ はデータ点である。

(4.38) を (4.52), (4.54), (4.55) を用いて書き換えると、

$$\begin{aligned}\hat{f}_j(x) = \sqrt{\frac{2}{h}}\Bigg(&\frac{\gamma_0}{n}\sum_{s=1}^{n} I_s\left(\frac{j-1}{N}, \frac{j}{N}\right) + \frac{\gamma_1}{n}\sum_{s=1}^{n}(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\\ &+ \frac{\gamma_2}{n}\sum_{s=1}^{n}(\mathrm{x}_s - t_j)^2 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\\ &+ \frac{\gamma_3}{n}\sum_{k=1}^{N-1} w_{j,k}\Bigg[-\frac{3}{2}\left\{\sum_{s=1}^{n} I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - \sum_{s=1}^{n} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\\ &- \frac{10}{h}\left\{\sum_{s=1}^{n}(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + \sum_{s=1}^{n}(\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\\ &+ \frac{42}{h^2}\left\{\sum_{s=1}^{n}(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - \sum_{s=1}^{n}(\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\Bigg]\\ &+ \frac{\gamma_4}{n}\sum_{k=1}^{N-1} w_{j-1,k}\Bigg[-\frac{3}{2}\left\{\sum_{s=1}^{n} I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - \sum_{s=1}^{n} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\\ &- \frac{10}{h}\left\{\sum_{s=1}^{n}(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + \sum_{s=1}^{n}(\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\\ &+ \frac{42}{h^2}\left\{\sum_{s=1}^{n}(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - \sum_{s=1}^{n}(\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\Bigg]\Bigg), \quad (4.56)\end{aligned}$$

ただし、$\gamma_0, \ldots, \gamma_4$ について $x - t_j \in \left[-\frac{h}{2},\ \frac{h}{2}\right)$ より、有限な値を取る。

(4.56) より、

$$\begin{aligned}Z_s = \frac{1}{n}\sqrt{\frac{2}{h}}\Bigg(&\gamma_0 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right) + \gamma_1(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\\ &+ \gamma_2(\mathrm{x}_s - t_j)^2 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\\ &+ \gamma_3\sum_{k=1}^{N-1} w_{j,k}\Bigg[-\frac{3}{2}\left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\\ &- \frac{10}{h}\left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\\ &+ \frac{42}{h^2}\left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\Bigg]\\ &+ \gamma_4\sum_{k=1}^{N-1} w_{j-1,k}\Bigg[-\frac{3}{2}\left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\\ &- \frac{10}{h}\left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\\ &+ \frac{42}{h^2}\left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\Bigg]\Bigg). \quad (4.57)\end{aligned}$$

このとき、

$$\hat{f}_j(x) \simeq \sum_{s=1}^{n} Z_s, \tag{4.58}$$

で、Z_s は独立同一分布に従う。

ここで、

$$E\left|Z_s - E\left[Z_s\right]\right|^3 \leq E\left|Z_s\right|^3 + 3E\left[Z_s^2\right]\left|E\left[Z_s\right]\right| + 4\left|E\left[Z_s\right]\right|^3. \tag{4.59}$$

Z_s^2 についてシュワルツの不等式より、

$$\begin{aligned}
Z_s^2 \leq \frac{18}{n^2 h}\Bigg[&\gamma_0^2\left\{I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2 + \gamma_1^2\left\{(\mathrm{x}_s - t_j)I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2 \\
&+ \gamma_2^2\left\{(\mathrm{x}_s - t_j)^2 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2 \\
&+ \frac{9\gamma_3^2}{4}\sum_{k=1}^{N-1} w_{j,k}^2\left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2 \\
&+ \frac{100\gamma_3^2}{h^2}\sum_{k=1}^{N-1} w_{j,k}^2\left\{(\mathrm{x}_s - t_{k+1})I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + (\mathrm{x}_s - t_k)I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2 \\
&+ \frac{1764\gamma_3^2}{h^4}\sum_{k=1}^{N-1} w_{j,k}^2\left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2 \\
&+ \frac{9\gamma_4^2}{4}\sum_{k=1}^{N-1} w_{j-1,k}^2\left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2 \\
&+ \frac{100\gamma_4^2}{h^2}\sum_{k=1}^{N-1} w_{j-1,k}^2\left\{(\mathrm{x}_s - t_{k+1})I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + (\mathrm{x}_s - t_k)I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2 \\
&+ \frac{1764\gamma_4^2}{h^4}\sum_{k=1}^{N-1} w_{j-1,k}^2\left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\Bigg].
\end{aligned} \tag{4.60}$$

(4.60) について期待値を取ると、

$$
\begin{aligned}
E\left[Z_s^2\right] \leq \frac{18}{n^2h}\Bigg(&\gamma_0^2 E\left[\left\{I_s\left(\frac{j-1}{N},\frac{j}{N}\right)\right\}^2\right] + \gamma_1^2 E\left[\left\{(\mathrm{x}_s - t_j)I_s\left(\frac{j-1}{N},\frac{j}{N}\right)\right\}^2\right] \\
&+ \gamma_2^2 E\left[\left\{(\mathrm{x}_s - t_j)^2 I_s\left(\frac{j-1}{N},\frac{j}{N}\right)\right\}^2\right] \\
&+ \frac{9\gamma_3^2}{4} E\left[\sum_{k=1}^{N-1} w_{j,k}^2\left\{I_s\left(\frac{k}{N},\frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N},\frac{k}{N}\right)\right\}^2\right] \\
&+ \frac{100\gamma_3^2}{h^2} E\left[\sum_{k=1}^{N-1} w_{j,k}^2\left\{(\mathrm{x}_s - t_{k+1})I_s\left(\frac{k}{N},\frac{k+1}{N}\right) + (\mathrm{x}_s - t_k)I_s\left(\frac{k-1}{N},\frac{k}{N}\right)\right\}^2\right] \\
&+ \frac{1764\gamma_3^2}{h^4} E\left[\sum_{k=1}^{N-1} w_{j,k}^2\left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N},\frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N},\frac{k}{N}\right)\right\}^2\right] \\
&+ \frac{9\gamma_4^2}{4} E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2\left\{I_s\left(\frac{k}{N},\frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N},\frac{k}{N}\right)\right\}^2\right] \\
&+ \frac{100\gamma_4^2}{h^2} E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2\left\{(\mathrm{x}_s - t_{k+1})I_s\left(\frac{k}{N},\frac{k+1}{N}\right) + (\mathrm{x}_s - t_k)I_s\left(\frac{k-1}{N},\frac{k}{N}\right)\right\}^2\right] \\
&+ \frac{1764\gamma_4^2}{h^4} E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2\left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N},\frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N},\frac{k}{N}\right)\right\}^2\right]\Bigg).
\end{aligned}
\tag{4.61}
$$

(4.61) の大括弧内の第 1 項目について、$I_s\left(\frac{k-1}{N},\frac{k}{N}\right) \overset{i.i.d.}{\sim} B\left(1,\ p_k\right),\ (k = 1, 2, \ldots, N)$ より、

$$
\begin{aligned}
E\left[\left\{I_s\left(\frac{j-1}{N},\frac{j}{N}\right)\right\}^2\right] &= \mathrm{Var}\left[I_s\left(\frac{j-1}{N},\frac{j}{N}\right)\right] + E\left[I_s\left(\frac{j-1}{N},\frac{j}{N}\right)\right]^2 \\
&= p_j\left(1 - p_j\right) + p_j^2 \\
&= p_j \\
&= \int_{B_j} f(t)dt \\
&\sim \int_{B_j} \left\{f(x) + (t-x)f'(x) + \cdots\right\} dt \\
&\sim hf(x).
\end{aligned}
\tag{4.62}
$$

したがって、前述の正則条件から $f^{(5)}(x) < \infty$ より、ある有限な定数 C_1 を用いて下記の関係が成り立つ;

$$
\gamma_0^2 E\left[\left\{I_s\left(\frac{j-1}{N},\frac{j}{N}\right)\right\}^2\right] < C_1. \tag{4.63}
$$

大括弧内の第 2 項目について、$\mathrm{x}_s - t_k \in \left[-\frac{h}{2}, \frac{h}{2}\right),\ (k = 1, 2, \ldots, N)$ を用いて、

$$\begin{aligned} E\left[\left\{(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] &= \mathrm{Var}\left[(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right] + E\left[(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right]^2 \\ &\leq h^2\left(\mathrm{Var}\left[I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right] + E\left[I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right]^2\right) \\ &= h^2\left\{p_j\left(1 - p_j\right) + p_j^2\right\} \\ &= h^2 p_j \\ &\sim h^3 f(x). \end{aligned} \tag{4.64}$$

したがって、同様にして、

$$\gamma_1^2 E\left[\left\{(\mathrm{x}_s - t_j) I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] < C_2. \tag{4.65}$$

大括弧内の第 3 項目について、

$$\begin{aligned} E\left[\left\{(\mathrm{x}_s - t_j)^2 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] &= \mathrm{Var}\left[(\mathrm{x}_s - t_j)^2 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right] + E\left[(\mathrm{x}_s - t_j)^2 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right]^2 \\ &\leq h^4\left(\mathrm{Var}\left[I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right] + E\left[I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right]^2\right) \\ &= h^4\left\{p_j\left(1 - p_j\right) + p_j^2\right\} \\ &= h^4 p_j \\ &\sim h^5 f(x). \end{aligned} \tag{4.66}$$

したがって、同様にして、

$$\gamma_2^2 E\left[\left\{(\mathrm{x}_s - t_j)^2 I_s\left(\frac{j-1}{N}, \frac{j}{N}\right)\right\}^2\right] < C_3. \tag{4.67}$$

大括弧内の第 4 項目について、

$$
\begin{aligned}
&E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] \\
&= \mathrm{Var}\left[\sum_{k=1}^{N-1} w_{j,k} \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\right] \\
&\qquad + E\left[\sum_{k=1}^{N-1} w_{j,k} \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}\right]^2 \\
&= \sum_{k=1}^{N-1} w_{j,k}^2 \left(\mathrm{Var}\left[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right)\right] + \mathrm{Var}\left[I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right]\right) \\
&\quad - 2\mathrm{Cov}\left(\sum_{k=1}^{N-1} w_{j,k} I_s\left(\frac{k}{N}, \frac{k+1}{N}\right), \sum_{k=1}^{N-1} w_{j,k} I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right) \\
&\quad + \sum_{k=1}^{N-1} w_{j,k}^2 \left(E\left[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right)\right]^2 + E\left[I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right]^2\right) \\
&\quad - 2\sum_{k=1}^{N-1} w_{j,k}^2 E\left[I_s\left(\frac{k}{N}, \frac{k+1}{N}\right)\right] E\left[I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right] \\
&= \sum_{k=1}^{N-1} w_{j,k}^2 \left\{p_{k+1}\left(1-p_{k+1}\right) + p_k\left(1-p_k\right)\right\} + 2\sum_{k=1}^{N-1} w_{j,k}^2 p_{k+1} p_k \\
&\quad + \sum_{k=1}^{N-1} w_{j,k}^2 \left(p_{k+1}^2 + p_k^2\right) - 2\sum_{k=1}^{N-1} w_{j,k}^2 p_{k+1} p_k \\
&= \sum_{k=1}^{N-1} w_{j,k}^2 \left(p_{k+1} + p_k\right) \\
&\sim 2\sum_{k=1}^{N-1} w_{j,k}^2 h f(x). \qquad (4.68)
\end{aligned}
$$

したがって、同様にして、

$$
\gamma_3^2 E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < C_4. \qquad (4.69)
$$

大括弧内の第 5 項目について、

$$
\begin{aligned}
&E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{ (\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] \\
&\leq h^2 E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{ I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right], \qquad (4.70)
\end{aligned}
$$

で表現され、(4.68) の通りに期待値部分が導出できることから、同様にして、

$$\left(\frac{100\gamma_3^2}{h^2}\right) E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < C_5. \tag{4.71}$$

大括弧内の第 6 項目について、

$$\begin{aligned} &E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] \\ &\quad \le h^4 E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right], \end{aligned} \tag{4.72}$$

で表現され、(4.68) の通りに期待値部分が導出できることから、同様にして、

$$\left(\frac{1764\gamma_3^2}{h^4}\right) E\left[\sum_{k=1}^{N-1} w_{j,k}^2 \left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < C_6. \tag{4.73}$$

大括弧内第 7〜9 項目は (4.69), (4.71), (4.73) と同様にして、

$$\gamma_4^2 E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2 \left\{I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < C_7, \tag{4.74}$$

$$\left(\frac{100\gamma_4^2}{h^2}\right) E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2 \left\{(\mathrm{x}_s - t_{k+1}) I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) + (\mathrm{x}_s - t_k) I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < C_8, \tag{4.75}$$

$$\left(\frac{1764\gamma_4^2}{h^4}\right) E\left[\sum_{k=1}^{N-1} w_{j-1,k}^2 \left\{(\mathrm{x}_s - t_{k+1})^2 I_s\left(\frac{k}{N}, \frac{k+1}{N}\right) - (\mathrm{x}_s - t_k)^2 I_s\left(\frac{k-1}{N}, \frac{k}{N}\right)\right\}^2\right] < C_9. \tag{4.76}$$

以上より、(4.63)〜(4.76) から $E\left[Z_s^2\right]$ は、

$$E\left[Z_s^2\right] < \frac{18}{n^2 h} \times (C_1 + C_2 + \cdots + C_9) \sim O\left(\frac{1}{n^2 h}\right). \tag{4.77}$$

また、$|Z_s| = O\left(\frac{1}{nh}\right)$ を用いて、

$$|Z_s|^3 < Z_s^2 O\left(\frac{1}{nh}\right), \tag{4.78}$$

であることから、

$$E\left|Z_s\right|^3 = O\left(\frac{1}{n^2 h}\right) \times O\left(\frac{1}{nh}\right) = O\left(\frac{1}{n^3 h^2}\right). \tag{4.79}$$

したがって、(4.59)、(4.77) より、

$$\sum_{s=1}^{n} E\left|Z_s - E\left[Z_s\right]\right|^3 = O\left(\frac{1}{n^2h^2}\right) + o\left(\frac{1}{n^2h}\right) + o\left(\frac{1}{n^2}\right)$$
$$= O\left(\frac{1}{n^2h^2}\right). \tag{4.80}$$

$\hat{f}_j(x)$ の分散を $\sigma\left[\hat{f}_j(x)\right]^2$ とすると、$\sigma\left[\hat{f}_j(x)\right]^2 \sim O\left(\frac{1}{nh}\right)$ より、

$$\frac{\sum_{s=1}^{n} E\left|Z_s - E\left[Z_s\right]\right|^3}{\sigma\left[\hat{f}_j(x)\right]^3} = \frac{O\left(\frac{1}{n^2h^2}\right)}{O\left(\frac{1}{n^{\frac{3}{2}}h^{\frac{3}{2}}}\right)} = O\left(\frac{1}{\sqrt{nh}}\right) = o\left(1\right). \tag{4.81}$$

5　Histogram のビン幅補正法

本章では、Histogram の推定区間と既知の閉区間の定義域を一致させるためのビン幅補正法について論じる。

5.1　ビン幅の補正法の構築

Histogram の定義域は既知の閉区間とする。この定義域について、一般性を失わず $[a,\ b]$ と表し、ただし、定数 $a,\ b$ は実数で、$a < b$ とする。定義域は既に決められているため、Histogram の始点を定義域の最小値 a とする。この条件のもとで、n 個の標本が与えられたとき、(2.4) より AMISE$[\hat{f}_{HIST}(x;h)]$ を最小にする理論的な最適ビン幅 h^* が決定される。したがって、h^* は標本から与えられた既知の実数値であるとして、ビン幅の補正を考える。h^* が決定されれば、定義域 $[a,\ b]$ におけるビンの最大個数は自動的に定まる。

このときのビンの最大個数を m とすると、

$$m = \frac{b-a}{h^*} + \frac{1}{2}, \tag{5.1}$$

で表現でき、m は h^* に依存する実数値の定数である。一般にビン数は整数値であるため、m を整数部分の m^* と小数部分の η に分解し、小数部分の η は剰余項として扱う。この時、m^* と η はそれぞれ以下の通り定義される;

$$m^* = \left[\frac{b-a}{h^*} + \frac{1}{2}\right], \tag{5.2}$$

$$\eta := [0,\ 1), \tag{5.3}$$

ただし、m^* はガウス記号を用いて表現している。

Histogram の推定区間と定義域とのずれを「ビン残差」と呼ぶことにし、ビン残差を δ とすると、δ は $\left[-\frac{h^*}{2},\ \frac{h^*}{2}\right]$ の値を取り、分布 f に依存する h^* との関係で決まることから、確率密度関数 $g(\delta)$ に従うものとする。このとき、δ は以下の通り表される;

$$\delta \sim g(\delta),\ \delta \in \left[-\frac{h^*}{2},\ \frac{h^*}{2}\right],$$

$$-\frac{h^*}{2} \leq E[\delta] \leq \frac{h^*}{2}, \tag{5.4}$$

$$0 \leq E[\delta^2] \leq \frac{h^{*2}}{4}, \tag{5.5}$$

ただし、$\int_{-h^*/2}^{h^*/2} g(x)dx = 1$ とする。

ビン残差 δ を各ビンに等配分してビン幅を補正することで、Histogram の推定区間と定義域 $[a,\ b]$ が一致する。したがって、ビン幅 h^* に対してビン残差 δ をビン数 m^* で等分配して補正す

る。この補正後のビン幅を $\tilde{h}$ とすると、

$$\tilde{h} = h^* - \frac{\delta}{m^*}, \tag{5.6}$$

となる。この時のビン B_j における $\hat{f}(x;\tilde{h})$ は以下の通り表現される;

$$\hat{f}(x;\tilde{h}) = \frac{\tilde{\nu}_j}{n\tilde{h}} = \frac{\tilde{\nu}_j}{n}\frac{m^*}{m^*h^* - \delta}, \tag{5.7}$$

ただし、$\tilde{\nu}_j$ はビン幅 $\tilde{h}$ を用いた時の B_j における度数である。

5.2 漸近的性質

本節ではビン幅補正後の Histogram について、MISE の意味で漸近一致性と漸近正規性が成り立つことを示す。

ビン幅補正後の Histogram に関して次の正則条件を満たすものとする;

(i) ビン幅 h について、$n \to \infty$ のとき、$h \to 0$ かつ $nh \to \infty$.
(ii) 関数 $f(x)$ は絶対連続関数で、導関数の一階積分が可能。

この (i) と (ii) の条件のもとで、以下の定理が成り立つ。ここで、h^* は AMISE$[\hat{f}_{HIST}(x)]$ を最小にする理論的な最適ビン幅、$\tilde{h}$ は補正後ビン幅を指すことに注意する。

Theorem 1：ビン幅補正後 Histogram の漸近一致性

ビン幅補正後の Histogram である $\hat{f}(x;\tilde{h})$ について、

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

が MISE の意味で漸近的に成り立つ。

Theorem 2：ビン幅補正後 Histogram の漸近正規性

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

$\alpha = \frac{1}{3}$ のとき、

$$\sqrt{nh^*}\left\{\hat{f}(x;\tilde{h}) - f(x)\right\} \xrightarrow{d} N\left(\mathrm{Bias}[\hat{f}(x;\tilde{h})],\ f(\xi_j)\left(1 + \frac{\delta}{m^*h^*}\right)\right), \tag{5.8}$$

$\alpha > \frac{1}{3}$ のとき、

$$\sqrt{nh^*}\left\{\hat{f}(x;\tilde{h}) - f(x)\right\} \xrightarrow{d} N\left(o(1),\ f(\xi_j)\left(1 + \frac{\delta}{m^*h^*}\right)\right), \tag{5.9}$$

が漸近的に成り立つ。ただし、$\xi_j \in B_j$ は平均値の定理 $p_j = \int_{B_j} f(t)dt = hf(\xi_j)$ 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 Histogram の漸近一致性の証明は 5.4 節の Appendix 1、漸近正規性の証明は 5.5 節の Appendix 2 で詳述する。

5.3 MISE の上限と下限

5.3.1 項では、補正後ビン幅を用いた Histogram の $E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right]$ の上限と下限を示す。また、一般的にビン残差の分布に関しては事前に情報を持たないため、一様分布に従うと仮定するのが自然だと考えられる。そのため、5.3.2 項ではビン残差 δ が一様分布に従う場合について取り扱う。

5.3.1 $E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right]$ の上限と下限

補正後のビン幅 $\tilde{h}$ における $\mathrm{AMISE}[\hat{f}(x;\tilde{h})]$ に対し、δ について期待値をとると、

$$E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] = E_\delta\left[n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] + E_\delta\left[\frac{R(f')}{12}h^{*2}\middle|h^*\right] - E_\delta\left[\frac{R(f')}{6}\frac{\delta}{m^*}h^*\middle|h^*\right] + E_\delta\left[\frac{R(f')}{12}\frac{\delta^2}{m^{*2}}\middle|h^*\right]. \tag{5.10}$$

(5.10) の第 1 項は $\frac{\delta}{m^*h^*} \sim O\left(n^{-\frac{1}{3}}\right)$ を用いて、$E\left[\frac{1}{nh^*}\middle|h^*\right] + E\left[\frac{1}{n}\frac{\delta}{m^*h^*}\middle|h^*\right]$ に分けることで、(5.10) で $E\left[\frac{1}{n}\frac{\delta}{m^*h^*}\middle|h^*\right] - E_\delta\left[\frac{R(f')}{6}\frac{\delta}{m^*}h^*\middle|h^*\right]$ は (2.4) から消し合う。

$E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right]$ は、2 次モーメントが最小となる $\delta = 0$ での一点分布の時に下限値を取り、2 次モーメントが最大となる $\delta = \pm\frac{h^*}{2}$ での左端一点分布、右端一点分布または両端二点分布の時に上限値を取る。したがって、$0 \le E[\delta^2] \le \frac{h^{*2}}{4}$ を (5.10) に代入して、

$$\frac{1}{nh^*} + \frac{R(f')}{12}h^{*2} \le E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \le \frac{1}{nh^*} + \frac{R(f')}{12}h^{*2} + \frac{R(f')}{48m^{*2}}h^{*2}. \tag{5.11}$$

$\mathrm{AMISE}[\hat{f}(x;h^*)] = \frac{1}{nh^*} + \frac{R(f')}{12}h^{*2}$ であるため、

$$\mathrm{AMISE}[\hat{f}(x;h^*)] \le E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \le \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{R(f')}{48m^{*2}}h^{*2}, \tag{5.12}$$

となる。

(5.12) について、$h^* \sim O\left(n^{-\frac{1}{3}}\right)$, $m^* \sim O\left(n^{\frac{1}{3}}\right)$ より、

$$\mathrm{AMISE}[\hat{f}(x;h^*)] \le E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \le \mathrm{AMISE}[\hat{f}(x;h^*)] + O(n^{-\frac{4}{3}}). \tag{5.13}$$

(5.13) から、ビン残差 δ の分布に関わらず、ビン幅の補正は Histogram の AMISE に影響しないことが示される。

5.3.2 ビン残差 δ が一様分布に従う場合

ビン残差 δ が一様分布に従う場合、$E[\delta^2] = \frac{h^{*2}}{12}$ であるため (5.10) に代入して、

$$E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] = \frac{1}{nh^*} + \frac{R(f')}{12}h^{*2} + \frac{R(f')}{144m^{*2}}h^{*2}. \tag{5.14}$$

$\mathrm{AMISE}[\hat{f}(x;h^*)] = \frac{1}{nh^*} + \frac{R(f')}{12}h^{*2}$ より、

$$E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] = \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{R(f')}{144m^{*2}}h^{*2}. \tag{5.15}$$

(5.15) について、$h^* \sim O\left(n^{-\frac{1}{3}}\right),\ m^* \sim O\left(n^{\frac{1}{3}}\right)$ より、

$$E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] = \mathrm{AMISE}[\hat{f}(x;h^*)] + O(n^{-\frac{4}{3}}). \tag{5.16}$$

以上より、ビン残差が一様分布に従う場合にも、ビン幅の補正は Histogram の推定精度に影響しない。

5.4 数値実験

有限標本における補正後ビン幅 $\tilde{h}$ が Histogram の推定精度に及ぼす影響を ISE の数値実験から調べる。標準正規分布 $N(0,1)$ について、ビン残差 δ が ISE に与える影響を調べるため、確率分布の右片側 tail 付近の確率が高い定義域 $[-3,\ 0]$ と右片側 tail 付近の確率が低い定義域 $[0,\ 3]$ について実験を行う。また、比較的定義域が広く、両側 tail 部分の確率が低い定義域 $[-3,\ 3]$ と、比較的定義域が狭く、両側 tail 部分の確率が高い定義域 $[-1,\ 1]$ について実験を行う。データ数は Histogram のビン数が同一にならないような $n = 50,\ 200,\ 500,\ 1000,\ 5000$ で、ビン幅はスコットのルールで推定し、ISE の数値実験 10000 回の平均 (MISE) と標準偏差を算出する。

5.4.1 定義域 $[-3,\ 0], [0,\ 3]$ での数値実験結果

表 5.1 は定義域 $[-3,\ 0]$ における補正なし最適ビン幅と補正後ビン幅の MISE、表 5.2 は定義域 $[0,\ 3]$ における補正なし最適ビン幅と補正後ビン幅の MISE の数値実験結果を示している。表は小さい値の方に下線が引いてある。ビン幅、定義域に関わらずデータ数が大きくなるにつれて MISE の値は小さくなる。

定義域 $[-3,\ 0]$ と $[0,\ 3]$ でどちらの場合も、データ数に関わらず補正後ビン幅の MISE が補正なし最適ビン幅の MISE より小さく、データが多い場合でもビン幅補正が有効であることが分かる。

定義域 $[-3,\ 0]$ と $[0,\ 3]$ の結果を比較すると、データ数に関わらず定義域 $[-3,\ 0]$ の方がビン幅補正による効果が大きい。このことから、分布の右片側 tail 部分の確率が大きい、すなわちビン残差部分にデータが多い場合に、ビン幅補正がより効果的であると考えられる。

表 5.1 定義域 $[-3,\ 0]$ の MISE

	$n = 50$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
補正なし最適ビン幅	0.071591	0.041823	0.029576	0.022289	0.008218
補正後ビン幅	<u>0.031194</u>	<u>0.014054</u>	<u>0.007924</u>	<u>0.005187</u>	<u>0.001882</u>

表 5.2　定義域 $[0,\ 3]$ の MISE

	$n = 50$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
補正なし最適ビン幅	0.032791	0.014271	0.008215	0.005302	0.001892
補正後ビン幅	0.031301	0.013852	0.008022	0.005208	0.001870

表 5.3 は定義域 $[-3,\ 0]$ の ISE の標準偏差、表 5.4 は定義域 $[0,\ 3]$ の ISE の標準偏差の数値実験結果を示している。表中で小さい値の方に下線が引いてある。定義域、ビン幅に関わらずデータ数が大きくなるにつれて ISE の標準偏差は小さくなっていくことが分かる。

定義域 $[-3,\ 0]$、定義域 $[0,\ 3]$ のどちらの場合も、補正後ビン幅の ISE の標準偏差の方が小さく、分散の安定化に有効であることが分かる。

表 5.3　定義域 $[-3,\ 0]$ の ISE 標準偏差

	$n = 50$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
補正なし最適ビン幅	0.060708	0.037314	0.027882	0.022329	0.010863
補正後ビン幅	0.021346	0.007496	0.003598	0.002119	0.000587

表 5.4　定義域 $[0,\ 3]$ の ISE 標準偏差

	$n = 50$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
補正なし最適ビン幅	0.022830	0.007671	0.003769	0.002200	0.000590
補正後ビン幅	0.020809	0.007383	0.003702	0.002136	0.000583

5.4.2　定義域 $[-3,\ 3]$ での数値実験結果

表 5.5 は補正なし最適ビン幅と補正後ビン幅の MISE、表 5.6 は補正なし最適ビン幅と補正後ビン幅の ISE の標準偏差の数値実験結果を示している。表は小さい値の方に下線が引いてある。ビン幅に関わらずデータ数が大きくなるにつれて MISE と ISE の標準偏差の値は小さくなる。補正なし最適ビン幅と補正後ビン幅で MISE と ISE の標準偏差をデータ数ごとに見ると、あまり差が見られない。そのため、両側 tail 部分の確率が低い場合には、ビン残差部分のデータが少なく、補正の効果が低いため優劣の判断はつかないと考えられる。一方で、ビン幅補正は分散安定化に有効である。

表 5.5　定義域 $[-3,\ 3]$ の MISE

	$n=50$	$n=200$	$n=500$	$n=1000$	$n=5000$
補正なし最適ビン幅	0.026028	0.011122	0.006210	0.004014	0.001417
補正後ビン幅	0.026370	0.011167	0.006200	0.004016	0.001416

表 5.6　定義域 $[-3,\ 3]$ の ISE 標準偏差

	$n=50$	$n=200$	$n=500$	$n=1000$	$n=5000$
補正なし最適ビン幅	0.013043	0.004352	0.002064	0.001182	0.000322
補正後ビン幅	0.012303	0.004156	0.002037	0.001171	0.000319

5.4.3　定義域 $[-1,\ 1]$ での数値実験結果

表 5.7 は補正なし最適ビン幅と補正後ビン幅の MISE、表 5.8 は補正なし最適ビン幅と補正後ビン幅の ISE の標準偏差の数値実験結果を示している。表は小さい値の方に下線が引いてある。ビン幅に関わらずデータ数が大きくなるにつれて MISE と ISE の標準偏差の値は小さくなる。補正なし最適ビン幅と補正後ビン幅で MISE と ISE の標準偏差を比較すると、データ数に関わらず補正後ビン幅での値の方が小さい。したがって、両側 tail 部分の確率が高い場合、すなわち、ビン残差部分にデータ数が多い場合、ビン幅補正は MISE の減少と分散安定化に有効である。

表 5.7　定義域 $[-1,\ 1]$ の MISE

	$n=50$	$n=200$	$n=500$	$n=1000$	$n=5000$
補正なし最適ビン幅	0.033822	0.020803	0.014906	0.007901	0.003599
補正後ビン幅	0.028330	0.013213	0.007626	0.004933	0.001822

表 5.8　定義域 $[-1,\ 1]$ の ISE 標準偏差

	$n=50$	$n=200$	$n=500$	$n=1000$	$n=5000$
補正なし最適ビン幅	0.027852	0.010154	0.006521	0.005376	0.001669
補正後ビン幅	0.023530	0.007802	0.003758	0.002153	0.000583

5.5　Appendix 1：$\hat{f}(x;\tilde{h})$ の漸近一致性の証明

漸近一致性の証明について、ビン幅補正後の推定量 $\hat{f}(x;\tilde{h})$ の AMISE から示す。ここでは、ビン残差について $\delta \sim U\left[-\frac{h^*}{2},\ \frac{h^*}{2}\right]$ のもとで証明する。$\mathrm{AMISE}[\hat{f}_{HIST}(x)] = \frac{1}{nh} + \frac{R(f')}{12}h^2$ に

$\hat{f}(x;\tilde{h})$ を代入して、

$$\begin{aligned}\mathrm{AMISE}[\hat{f}(x;\tilde{h})] &= \frac{1}{n\tilde{h}} + \frac{1}{12}\tilde{h}^2 R(f') \\ &= \frac{1}{n\left(h^* - \frac{\delta}{m}\right)} + \frac{1}{12}\left(h^* - \frac{\delta}{m}\right)^2 R(f') \\ &= n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1} + \frac{R(f')}{12}h^* - \frac{R(f')}{6}\frac{\delta}{m^*}h^* + \frac{R(f')}{12}\frac{\delta^2}{m^{*2}}. \end{aligned} \tag{5.17}$$

ここで、$\mathrm{AMISE}[\hat{f}(x;\tilde{h})]$ に対し、δ について期待値をとると、

$$\begin{aligned}E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] = E_\delta\left[n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] + E_\delta\left[\frac{R(f')}{12}h^{*2}\middle|h^*\right] \\ - E_\delta\left[\frac{R(f')}{6}\frac{\delta}{m^*}h^*\middle|h^*\right] + E_\delta\left[\frac{R(f')}{12}\frac{\delta^2}{m^{*2}}\middle|h^*\right]. \end{aligned} \tag{5.18}$$

(5.18) の第 1 項について、

$$\begin{aligned}E_\delta\left[n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] &= n^{-1}E_\delta\left[\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] \\ &= \frac{m^*}{n}E_\delta\left[\frac{1}{m^*h^* - \delta}\middle|h^*\right] \\ &= \frac{m^*}{n}\int_{-\frac{h^*}{2}}^{\frac{h^*}{2}} \frac{1}{m^*h^* - \delta}\frac{1}{h^*}d\delta \\ &= \frac{m^*}{n}\frac{1}{h^*}\left[\log|m^*h^* - \delta|\right]_{-\frac{h^*}{2}}^{\frac{h^*}{2}} \\ &= \frac{m^*}{n}\frac{1}{h^*}\log\left(\frac{m^*h^* + \frac{h^*}{2}}{m^*h^* - \frac{h^*}{2}}\right) \\ &= \frac{m^*}{n}\frac{1}{h^*}\log\left(\frac{1 + \frac{1}{2m^*}}{1 - \frac{1}{2m^*}}\right). \end{aligned} \tag{5.19}$$

(5.19) の $\log\left(1 + \frac{1}{m^*}\right)$ の部分について、対数の原点まわりの級数展開により、

$$\log\left(\frac{1 + \frac{1}{2m^*}}{1 - \frac{1}{2m^*}}\right) \sim 2\left\{\frac{1}{2m^*} + \frac{1}{3}\left(\frac{1}{2m^*}\right)^3 + \frac{1}{5}\left(\frac{1}{2m^*}\right)^5 + \cdots\right\},$$

と近似される。

したがって、(5.18) の第 1 項は以下の式で近似される;

$$E_\delta\left[n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] \sim \frac{1}{nh^*}\left(1 + \frac{1}{3}\left(\frac{1}{2m^*}\right)^2\right). \tag{5.20}$$

(5.18) の第 2 項は、δ を含まない項のため定数となり、

$$E_\delta\left[\frac{R(f')}{12}h^{*2}\middle|h^*\right] = \frac{R(f')}{12}h^{*2}. \tag{5.21}$$

(5.18) の第 3 項について、$E_\delta[\delta] = 0$ より、

$$E_\delta\left[\frac{R(f')}{6}\frac{\delta}{m^*}h^*\middle|h^*\right] = 0. \tag{5.22}$$

(5.18) の第 4 項について、$E_\delta[\delta^2] = \frac{h^{*2}}{12}$ より、

$$E_\delta\left[\frac{R(f')}{12}\frac{\delta^2}{m^{*2}}\middle|h^*\right] = \frac{R(f')}{12m^{*2}}E_\delta[\delta^2] = \left(\frac{h^*}{12m^*}\right)^2 R(f'). \tag{5.23}$$

以上 (5.20)〜(5.23) 式より、$E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right]$ は、以下の通りである;

$$\begin{aligned} E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right] &\sim \frac{1}{nh^*} + \frac{1}{nh^*}\left(\frac{1}{12m^{*2}}\right) + \frac{R(f')}{12}h^{*2} + \frac{1}{12m^{*2}}\frac{R(f')}{12}h^{*2} \\ &= \left(1 + \frac{1}{12m^{*2}}\right)\mathrm{AMISE}[\hat{f}(x;h^*)]. \end{aligned} \tag{5.24}$$

(5.24) で $\mathrm{AMISE}[\hat{f}(x;h^*)]$ の係数部分 $\frac{1}{12m^{*2}}$ について、ビン数 $m^* = m - \eta,\ m = \frac{b-a}{h^*} + \frac{1}{2}$ 及び、ビン幅 $h^* = \left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}$ を代入して、

$$\begin{aligned} \frac{1}{12m^{*2}} &= \frac{1}{12}\left\{\frac{b-a}{h^*} - \left(\eta - \frac{1}{2}\right)\right\}^{-2} \\ &= \frac{h^{*2}}{12\left\{(b-a) - \left(\eta - \frac{1}{2}\right)h^*\right\}^2} \\ &= \frac{\left(\frac{6}{R(f')}\right)^{\frac{2}{3}} n^{-\frac{2}{3}}}{12\left\{(b-a) - \left(\eta - \frac{1}{2}\right)\left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}\right\}^2}, \end{aligned} \tag{5.25}$$

となる。

(5.25) を用いると、$E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right]$ は次の通りである;

$$\begin{aligned} E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right] &= \left(1 + \frac{1}{12m^{*2}}\right)\mathrm{AMISE}[\hat{f}(x;h^*)] \\ &= \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{\left(\frac{6}{R(f')}\right)^{\frac{2}{3}} n^{-\frac{2}{3}}}{12\left\{(b-a) - \left(\eta - \frac{1}{2}\right)\left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}\right\}^2}\mathrm{AMISE}[\hat{f}(x;h^*)]. \end{aligned} \tag{5.26}$$

(5.26) の第 2 項で $\mathrm{AMISE}[\hat{f}(x;h^*)] = \frac{3}{2}\left(\frac{R(f')}{6}\right)^{\frac{1}{3}} n^{-\frac{2}{3}}$ を用いて整理すると、

$$E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})\right]] = \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{\left(\frac{6}{R(f')}\right)^{\frac{1}{3}}}{8\left\{(b-a) - \left(\eta - \frac{1}{2}\right)\left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}\right\}^2} n^{-\frac{4}{3}}. \tag{5.27}$$

補正後のビン幅 $\tilde{h}$ に基づく Histogram の AMISE$[\hat{f}(x;\tilde{h})]$ の期待値は、最小 AMISE$[\hat{f}(x;h^*)]$ に (5.27) の第 2 項を加えたものとなる。また、(5.27) の第 2 項の分母に $b-a$ が含まれることから、定義域 $[a,\ b]$ が広くなるほど、$E\left[\text{AMISE}[\hat{f}(x;\tilde{h})]\right]$ は AMISE$[\hat{f}(x;h^*)]$ に近づいていく。

以上をまとめると、(5.27) で $h^* \propto n^{-\frac{1}{3}},\ \text{AMISE}[\hat{f}(x;h^*)] = O\left(n^{-\frac{2}{3}}\right)$ より、

$$\begin{aligned} E\left[\text{AMISE}[\hat{f}(x;\tilde{h})]\right] &= \text{AMISE}[\hat{f}(x;h^*)] + O\left(n^{-\frac{4}{3}}\right) \\ &= \text{AMISE}[\hat{f}(x;h^*)] + o\left(n^{-\frac{2}{3}}\right), \end{aligned} \tag{5.28}$$

であることから、$n \to \infty$ のとき、

$$E\left[\text{AMISE}[\hat{f}(x;\tilde{h})]\right] \xrightarrow{d} \text{AMISE}[\hat{f}(x;h^*)], \tag{5.29}$$

となり、

$$\int \left(\hat{f}(x;\tilde{h}) - f(x)\right)^2 dx \xrightarrow{d} 0, \tag{5.30}$$

であることから、

$$\hat{f}(x;\tilde{h}) \xrightarrow{d} f(x), \tag{5.31}$$

が成り立つ。以上から $\hat{f}(x;\tilde{h})$ が MISE の意味で $f(x)$ に漸近的に一致することが示された。

5.6 Appendix 2：$\hat{f}(x;\tilde{h})$ の漸近正規性の証明

ビン幅補正後の推定量 $\hat{f}(x;\tilde{h})$ の漸近正規性について示す。ビン B_j における $\hat{f}(x;\tilde{h})$ は以下の通りである;

$$\hat{f}(x;\tilde{h}) = \frac{\tilde{\nu}_j}{n\tilde{h}} = \frac{\tilde{\nu}_j}{n}\frac{m^*}{m^*h^* - \delta}. \tag{5.32}$$

$\tilde{\nu}_j \sim B\left(n,\ \tilde{p}_j\right)$ で、$\tilde{p}_j = \int_{B_j} f(t)dt$ とすると、$\hat{f}(x;\tilde{h})$ の期待値及び分散は、

$$E\left[\hat{f}(x;\tilde{h})\right] = \frac{m^*}{n\left(m^*h^* - \delta\right)} E\left[\tilde{\nu}_j\right] = \frac{m^*}{m^*h^* - \delta}\tilde{p}_j, \tag{5.33}$$

$$\begin{aligned} \text{Var}\left[\hat{f}(x;\tilde{h})\right] &= \frac{1}{\left(n\tilde{h}\right)^2}\text{Var}\left[\tilde{\nu}_j\right] \\ &= \frac{\tilde{p}_j\left(1-\tilde{p}_j\right)}{n\left(h^* - \frac{\delta}{m^*}\right)^2}. \end{aligned} \tag{5.34}$$

$\tilde{p}_k$ について積分の平均値の定理より、

$$\tilde{p}_j = \int_{B_j} f(t)dt = \tilde{h}f(\xi_j) = \left(h^* - \frac{\delta}{m^*}\right) f(\xi_j), \tag{5.35}$$

ただし、ξ_j は $\xi_j \in B_j$ を満たす B_j 内のある点とする。

(5.34) に (5.35) を代入すると、

$$\begin{aligned}
\mathrm{Var}\left[\hat{f}(x;\tilde{h})\right] &= \frac{\left(h^* - \frac{\delta}{m^*}\right) f(\xi_j) \left\{1 - \left(h^* - \frac{\delta}{m^*}\right) f(\xi_j)\right\}}{n\left(h^* - \frac{\delta}{m^*}\right)^2} \\
&= \frac{f(\xi_j)}{n\left(h^* - \frac{\delta}{m^*}\right)} + O\left(n^{-1}\right) \\
&\sim \frac{m^* f(\xi_j)}{n} \frac{1}{m^* h^* - \delta} \\
&\sim \frac{f(\xi_j)}{nh^*}\left(1 + \frac{\delta}{m^* h^*}\right). \qquad (5.36)
\end{aligned}$$

したがって、二項分布の中心極限定理と (5.36) を用いることで、次の通りの表現を得る。

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

$\alpha = \frac{1}{3}$ のとき、

$$\sqrt{nh^*}\left\{\hat{f}(x;\tilde{h}) - f(x)\right\} \xrightarrow{d} N\left(\mathrm{Bias}[\hat{f}(x;\tilde{h})],\ f(\xi_j)\left(1 + \frac{\delta}{m^* h^*}\right)\right), \qquad (5.37)$$

$\alpha > \frac{1}{3}$ のとき、

$$\sqrt{nh^*}\left\{\hat{f}(x;\tilde{h}) - f(x)\right\} \xrightarrow{d} N\left(o(1),\ f(\xi_j)\left(1 + \frac{\delta}{m^* h^*}\right)\right), \qquad (5.38)$$

が漸近的に成り立つ。以上より、$\hat{f}(x;\tilde{h})$ は漸近正規性が成り立つことが証明された。

6 Random Partitioned Histogram

本章では、ランダムな分割点による不等間隔 Histogram を繰り返し推定し、その平均を推定量とする Random Partitioned Histogram(以降、RPH) について論じる。

6.1 推定量の構築

RPH 推定の手順は大きく以下の 3 つに分けられる。

(i) ビン数を任意またはビン数とビン幅推定のルールに基づいて決定する。
(ii) 定義域を与え、ビン数に対応した区間を一様乱数から決定して不等間隔 Histogram を構築する。
(iii) 手順 (ii) の操作を繰り返し、推定した Histogram の平均を RPH 推定量とする。

手順 (i) では、分析者の任意または、ビン数の決定法としてよく用いられるスタージェスのルールやスコットのルール等のビン幅推定法を用いて Histogram の分割数を決定する。

手順 (ii) で、一般に Histogram の始点と終点は未知であるが、閉区間の定義域が与えられた場合の推定とし、定義域の最小値を始点、最大値を終点とする。手順 (i) で決定したビン数に対応した区間を一様乱数で決定する。このランダムに決定された分割点を用いて不等間隔 Histogram の推定を行う。図 6.1 は，繰り返しなしのランダムに分割したビンでの Histogram 推定の例である。定義域 $[-4,\ 4]$ で標準正規分布に従う 50 個のデータで、ビン数を 8 として乱数で決定した分割点に基づく Histogram の例を示す。データは固定しているが、分割点が毎回ランダムに決定されるため、異なる Histogram が推定される。また、一般的な不等間隔 Histogram と同様に、確率密度の条件を満たすように各区間のビン幅に比例して高さを調節している。

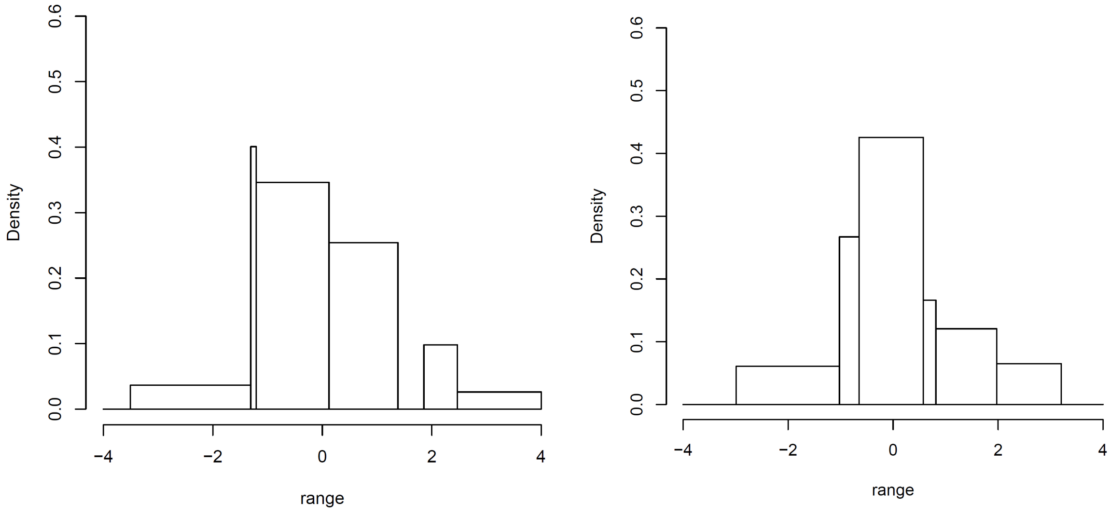

図 6.1　Random Partition の例 (ビン数 8、繰り返しなし)

手順 (iii) では、手順 (ii) を繰り返して複数の Histogram を構築し、その平均を RPH 推定量とする。図 6.2 の実線のグラフは、標準正規分布に従う 200 個のデータについて、任意のビン数 15 で繰り返し回数 50 の RPH 推定量である。RPH は繰り返し発生させた複数の Histogram の平均であるため、平滑化する特徴を持つ。したがって、データがあまり集中しない範囲では推定量を上方へ押し上げ、反対にデータが集中する範囲では推定量を下方へ引っ張る傾向がある。この傾向を軽減させる簡単な方法としては、定義域をデータの範囲よりも十分広く設定することである。したがって、定義域が与えられた RPH について数値実験を行うが、その際には広めの定義域を設定する。

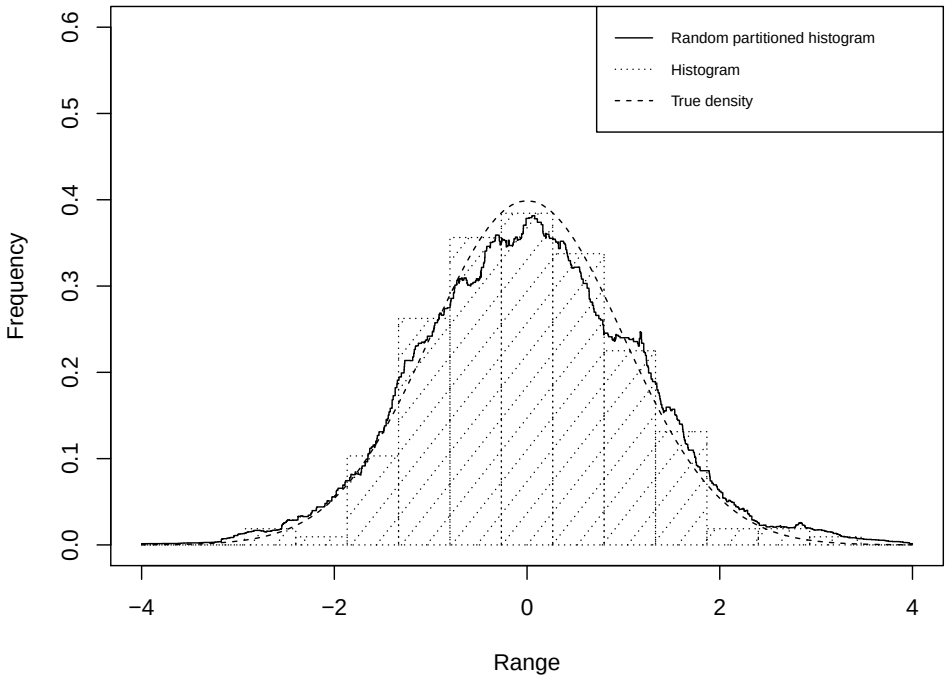

図 6.2　RPH 推定量の例

6.2　単峰の分布における数値実験設定

Histogram と RPH の密度推定の精度を比較するため、ISE について数値実験を行う。ここでは、MISE の変動を評価するため、ISE の数値実験を 10000 回行い、ISE 値とその標準偏差を計算する。以降、数値実験 10000 回の ISE の平均値を「ISE 値」と呼ぶ。定義域 $[-4,\ 4]$ で標準正規分布 $N(0,\ 1)$ に従うデータについて、表 6.1 で示す通りの設定で数値実験を行う。表 6.1 でデータ数を n、ビン数を $m_{(RPH)}$、繰り返し回数を r とし、以降も同様とする。

表 6.1　各数値実験における設定

	データ数 n	ビン数 $m_{(RPH)}$	繰り返し回数 r
ケース①	50, 100, 200, 500, 1000, 5000	$m_{(RPH)} = Scott$	30
ケース②	50	3, $m_{(RPH)} = Scott$, 13	2, 5, 10, 20, 50, 100
ケース③	50, 100, 200, 500, 1000, 5000	$m_{(RPH)} = Scott$	2, 3, 4, $\cdots$, 20, 50, 100
ケース④	50, 100, 200, 500, 1000, 5000	3, 5, $\cdots$, 47, 50	30

全ての数値実験において、Histogram の始点を定義域の最小値、終点を定義域の最大値とする。

以下、各数値実験の詳細を説明する。

ケース①：データ数を変化させたときの Histogram と RPH との推定精度の比較を目的とする。$n = 50, 100, 200, 500, 1000, 5000$ とする。Histogram については、スコットのルールで最適ビン幅を推定し、定義域を最適ビン幅で割った時の整数部分をビン数とする。以降、最適ビン数を用いる時は $m_{(RPH)} = Scott$ と記す。RPH の設定では、$m_{(RPH)}$ をスコットのルールで算出し、$m_{(RPH)} - 1$ 個の分割点を一様分布 $U(-4,\ 4)$ に従う乱数で決定し、その分割点に従って不等間隔 Histogram を構築する。繰り返し上記の Histogram を計算し、その平均を取って RPH を推定し、ISE を計算する。$r = 30$ で固定する。

ケース②：データ数を固定し、ビン数と繰り返し回数の変化による RPH の推定精度を調べることを目的とする。ここで、$n = 50$ に固定する。$m_{(RPH)}$ に関して、スコットのルールによる Histogram 推定を 10000 回繰り返した時の最適なビン数は平均 8.15 であった。したがって、最適なビン数約 8 より少ないビン数を 3、多いビン数を 13 と設定する。RPH を $m_{(RPH)} = Scott, m_{(RPH)} = 3, m_{(RPH)} = 13$ でそれぞれ推定し、ISE の計算を行う。$r = 2, 5, 10, 20, 50, 100$ とする。

ケース③： RPH 推定におけるデータ数と繰り返し回数を変化させた時の推定精度への影響を調べることを目的とする。ここではデータ数ごとに、各繰り返し回数における RPH の ISE 値を計算する。$n = 50, 100, 200, 500, 1000, 5000$ とし、$m_{(RPH)} = Scott$ で推定する。また、$r = 2, 3, 4, \cdots, 20, 50, 100$ とする。

ケース④： データ数とビン数を変化させた時の RPH の推定精度を調べることを目的とする。ここで、$r = 30$ で固定し、$n = 50, 100, 200, 500, 1000, 5000$ と、$m_{(RPH)} = 3, 5, \cdots, 47, 50$ とし、各データ数及びビン数における ISE 値を計算する。加えて、$m_{(RPH)}$ を変化させた時の ISE 値を調べる。また、RPH の ISE 値が、$n = 200,\ 500$ における Histogram の ISE 値を下回る $m_{(RPH)}$ の範囲について調べる。

6.3 単峰の分布における数値実験結果

(i) ケース①の結果

表 6.2 は、データ数を変化させた時の最適ビン数に基づく ISE 値の計算結果を示す。表で、比較して値が小さい方に下線を引いてある。Histogram と RPH のどちらも、n が増加するにつれて ISE 値は小さくなる。n に関わらず、RPH の方が ISE 値が小さい。

表 6.2 ケース①：ISE 値 ($m_{(RPH)} = Scott$)

	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
Histogram	0.026160	0.017032	0.011116	0.006284	0.003999	0.001410
RPH($r = 30$)	<u>0.018091</u>	<u>0.010741</u>	<u>0.006251</u>	<u>0.003016</u>	<u>0.001748</u>	<u>0.000528</u>

表 6.3 は、データ数を変化させた時の、推定値の安定性を表す ISE の標準偏差の計算結果を示す。ただし、表で、小さい値の方に下線を引いてある。Histogram と RPH のどちらも、n が増加するにつれて ISE の標準偏差は小さくなる。n に関わらず、RPH の方が ISE の標準偏差は小さい。Histogram と比較して、RPH の方が安定化することが分かる。

表 6.3　ケース①：ISE 標準偏差 ($m_{(RPH)} = Scott$)

	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
Histogram	0.012104	0.007022	0.004168	0.002071	0.001166	0.000312
RPH($r = 30$)	0.009641	0.005752	0.003220	0.001444	0.000769	0.000189

(ii) ケース②の結果

表 6.4 は、繰り返し回数を変化させた時の ISE 値の計算結果を示す。表で、$n = 50$ における Histogram の ISE 値 0.026160 を下回る箇所に下線を引いてある。$m_{(RPH)} = 3, m_{(RPH)} = Scott, m_{(RPH)} = 13$ のどの場合でも、r が増加するにつれて ISE 値は小さくなる。r に関わらず $m_{(RPH)} = 13$ の場合、ISE 値が相対的に最も小さい。Histogram の ISE 値と比較して、$m_{(RPH)} = Scott, 13$ の場合では、$r \geq 5$ で ISE 値が小さくなる。一方で、$m_{(RPH)} = 3$ の場合、r を増やしても Histogram より推定精度は劣る。したがって、最適ビン数を超える $m_{(RPH)} = 13$ で相対的に最も推定精度が優れることから、RPH のビン数は Histogram の最適ビン数より多く設定する方が良いと考えられる。

表 6.4　ケース②：ISE 値 ($n = 50$)

	$r = 2$	$r = 5$	$r = 10$	$r = 20$	$r = 50$	$r = 100$
Histogram	0.026160					
$m_{(RPH)} = 3$	0.097327	0.087727	0.084482	0.083044	0.081971	0.081602
$m_{(RPH)} = Scott$	0.035597	0.024235	0.020478	0.018602	0.017591	0.017244
$m_{(RPH)} = 13$	0.028028	0.018019	0.014704	0.012856	0.011889	0.011572

表 6.5 は、繰り返し回数を変化させた時の ISE の標準偏差の計算結果を示す。表で、$n = 50$ における Histogram の ISE の標準偏差 0.012104 を下回る箇所に下線を引いてある。$m_{(RPH)} = 3, m_{(RPH)} = Scott, m_{(RPH)} = 13$ のどの場合も、r が増加するにつれて ISE の標準偏差は小さくなる。r に関わらず $m_{(RPH)} = 13$ の場合、ISE の標準偏差が相対的に最も小さい。ケース①の結果から、Histogram と比較して、$m_{(RPH)} = Scott, m_{(RPH)} = 13$ の場合は $r \geq 5$、$m_{(RPH)} = 3$ の場合は $r \geq 20$ で標準偏差がより小さくなる。分散については、繰り返し回数に依存するものの、ビン数が $m_{(RPH)} = Scott$ 以上の場合だけでなく、少ない場合であっても Histogram より安定化が図られる。

表 6.5　ケース②：ISE 標準偏差 ($n = 50$)

	$r = 2$	$r = 5$	$r = 10$	$r = 20$	$r = 50$	$r = 100$
Histogram	0.012104					
$m_{(RPH)} = 3$	0.026456	0.018560	0.014095	0.010336	0.007370	0.006080
$m_{(RPH)} = Scott$	0.017645	0.012062	0.010489	0.009986	0.009692	0.009471
$m_{(RPH)} = 13$	0.013654	0.008749	0.007771	0.007420	0.007110	0.007158

(iii) ケース③の結果

図 6.3 は、異なるデータ数ごとの各繰り返し回数を変化させた時の ISE 値である。グラフ内で、曲線がなだらかになる箇所の一番左側を点で示している。n に関わらず、$2 \leq r \leq 10$ の時に ISE 値が大きく減少し、$r \geq 20$ ではほぼ変化が見られない。そのため、r をあまり大きくする必要はないことが分かる。したがって、曲線でなだらかになり始める一番左側の箇所を繰り返し回数に選択することを推奨する。この選択は、エルボー法と類似するものであり、エルボー法はクラスタリング手法の一つである k-means 法でクラスター数決定の際に良く用いられる。図 6.3 におけるグラフの形状がエルボー法で用いられるエルボーカーブに似ているため、それを適用した手法である。

データ数ごとの曲線に着目すると、$n = 50$ の場合には、$10 \leq r \leq 20$ で ISE 値が大きく減少する一方で、$n = 1000, 5000$ の場合には $5 \leq r \leq 10$ で ISE 値の減少幅が非常に小さくなる。このことから、データ数が小さい場合には繰り返し回数を多くした方が良い推定が得られる。一方で、データ数が大きい場合には、少ない繰り返し回数であっても十分精度の高い推定となる。

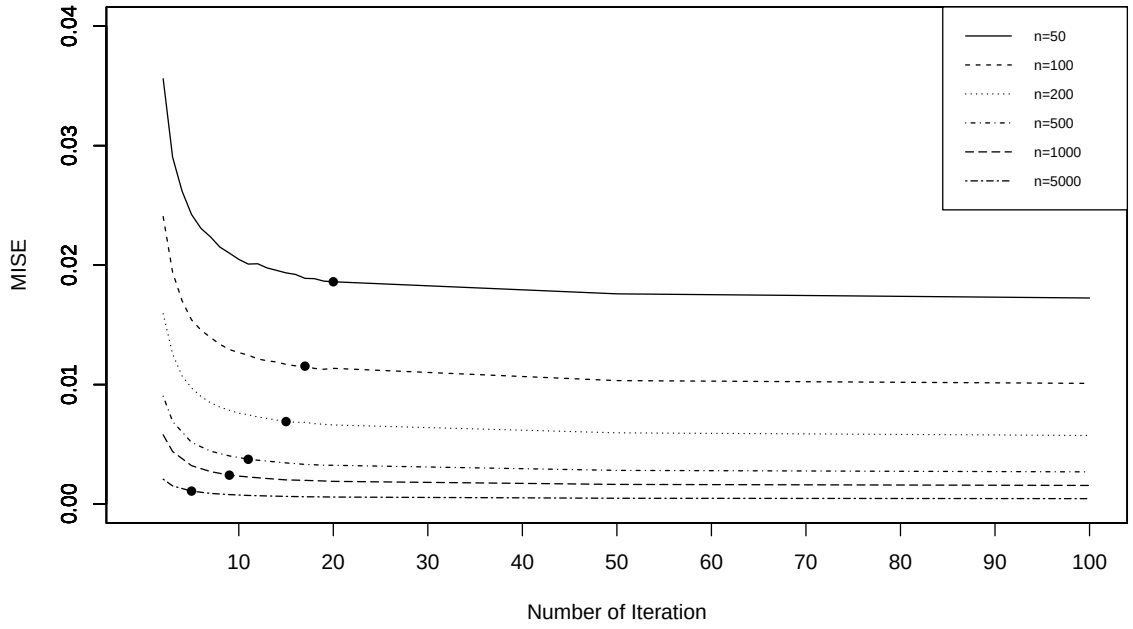

図 6.3　各繰り返し回数を変化させた時のデータ数別の ISE 値 ($m_{(RPH)} = Scott$) (黒点はエルボーカーブによる繰り返し回数の選択点を示している。)

(iv) ケース④の結果

図 6.4 は各データ数を固定し、ビン数を変化させた時の ISE 値である。n に関わらず、$m_{(RPH)}$ によって Histogram よりも推定精度が良くなることが明らかになった。しかしながら、n が大きくなるにつれて改良の程度は小さい。また、どの n においても、RPH の ISE 値が最小となる $m_{(RPH)}$ は Histogram の最適な $m_{(RPH)} = Scott$ より大きい。

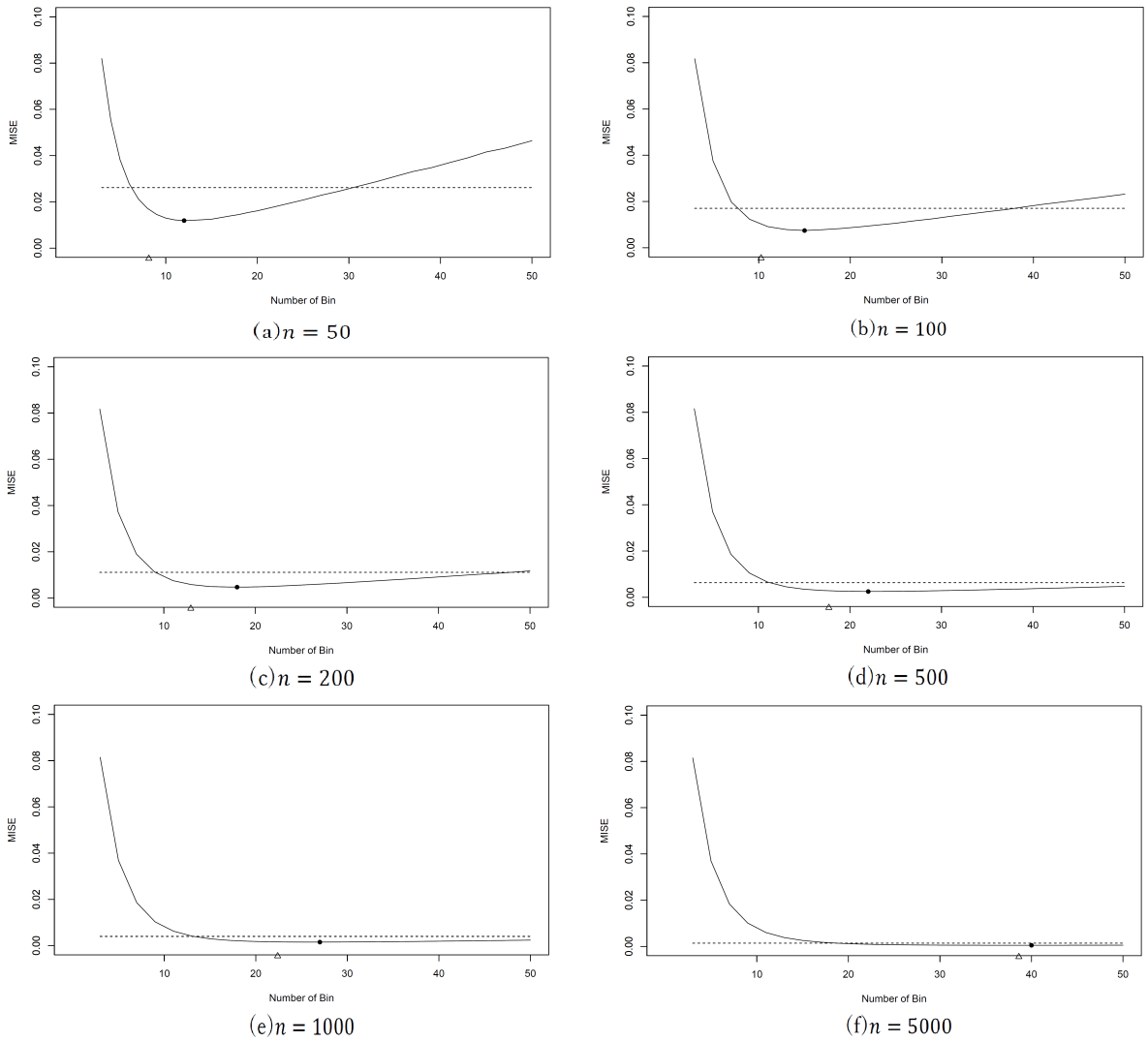

図 6.4 データ数ごとの各ビン数における ISE 値 ($r = 30$) (実線 : RPH、点線 : Histogram($m_{(RPH)} = Scott$)、丸印 : ISE 値の最小値、三角印 : スコットのルールによる平均ビン数)

図 6.5 は $n = 200$ の Histogram と、$n = 100, 200$ の RPH の ISE 値を比較したグラフである。グラフ内で、Histogram の $n = 200$ における ISE 値は 0.011116 である。Histogram の ISE 値を RPH が下回っているのは、$n = 100$ では $10 \leq m_{(RPH)} \leq 25$、$n = 200$ では $9 \leq m_{(RPH)} \leq 48$ である。

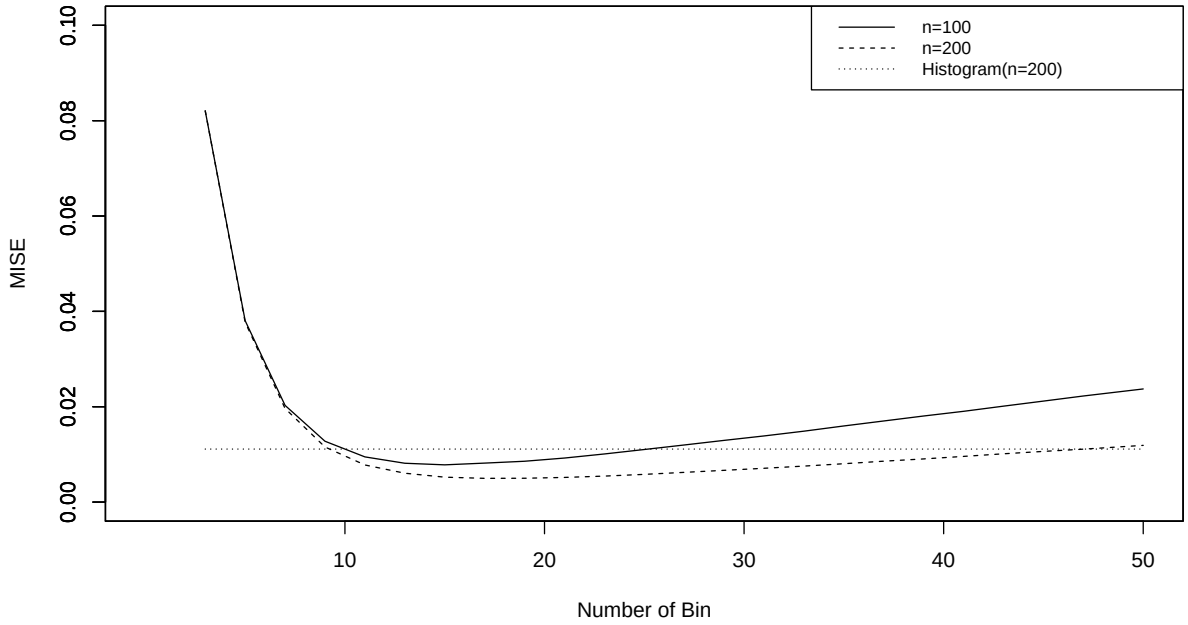

図 6.5　Histogram($n = 200, m_{(RPH)} = Scott$) と RPH($r = 30$) の ISE 値の比較

図 6.6 は $n = 500$ の Histogram と、$n = 200, 500$ の RPH の ISE 値を比較したグラフである。グラフ内で、Histogram の $n = 500$ における ISE 値は 0.006284 である。Histogram の ISE 値を RPH が下回っているのは、$n = 200$ では $12 \leq m_{(RPH)} \leq 27$、$n = 500$ では $m_{(RPH)} \geq 11$ である。

図 6.5、図 6.6 から、Histogram より少ないデータ数でも RPH の推定精度が良い場合があった。そのため、RPH のビン数を適切に選択すればデータ数が少なくても Histogram よりも優れた推定が可能である。

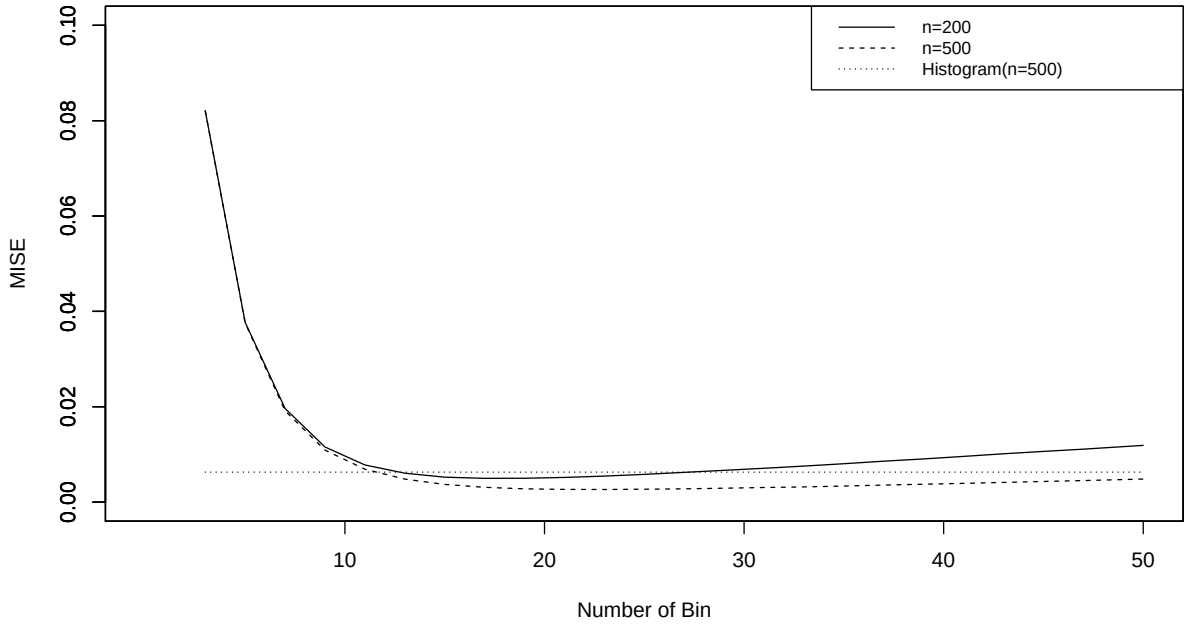

図 6.6　Histogram($n = 500, m_{(RPH)} = Scott$) と RPH($r = 30$) の ISE 値の比較

表 6.6 は、数値実験で得られたデータ数ごとの RPH の ISE 値の最小値及びその時のビン数、Histogram の数値実験における平均ビン数、RPH の ISE 値が最小時のビン数と Histogram の平均ビン数の比率である。ただし、表中の mh は Histogram における平均ビン数とする。Histogram と RPH どちらもビン数の推定にはスコットのルールを用いている。n に関わらず、RPH の ISE 値が最小となる $m_{(RPH)}$ は mh より多いことが分かった。また、今回選択した n における $m_{(RPH)}/mh$ の平均は 1.31 であった。加えて上記で示した通り、RPH では Histogram よりも多いビン数を選択する方が推定精度は良い傾向がある。したがって、RPH の最適ビン数の選択法について、単峰の分布の場合、ビン数の目安としては Histogram の 1.5 倍を推奨する。

表 6.6　各データ数における ISE 値の最小値及びその時のビン数 ($r = 30$)

	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
最小 ISE 値	0.012397	0.007809	0.004895	0.002638	0.001624	0.000525
$m_{(RPH)}$	12	15	18	23	27	40
mh	8.15	10.25	12.94	17.70	22.39	38.62
$m_{(RPH)}/mh$	1.47	1.46	1.39	1.30	1.21	1.04

6.4　多峰の分布における数値実験

ここまで真の分布が単峰の場合について見てきたが、多峰の分布における RPH の性質を確かめるために、混合正規分布における Histogram と RPH の推定精度を比較する。定義域 $[-6,\ 6]$ の混合正規分布 $\frac{3}{4}N\left(-1,\ \left(\frac{3}{2}\right)^2\right)+\frac{1}{4}N\left(2,\ \left(\frac{1}{3}\right)^2\right)$ に従うデータについて、Histogram と RPH それぞれの ISE の数値実験を 10000 回行い、その結果を比較する。$n=50,100,200,500,1000,5000$、$r=30$ とする。Histogram のビン数はスコットのルールで推定する。一方で、RPH の $m_{(RPH)}$ については 6.3 節の結果から、Histogram より多い方が望ましいことが明らかである。そのため、$m_{(RPH)}$ は mh を 1.5 倍した値の整数部分を使用する。表 6.7 は各データ数における Histogram の平均ビン数と、RPH での選択ビン数である。また、図 6.7 は $n=500$ における RPH 推定量の例である。

表 6.7　Histogram の平均ビン数及び RPH で選択したビン数

	$n=50$	$n=100$	$n=200$	$n=500$	$n=1000$	$n=5000$
mh	6.5	8.2	10.4	14.3	18.1	31.2
$m_{(RPH)}$ $(=mh\times 1.5)$	10	12	16	21	27	47

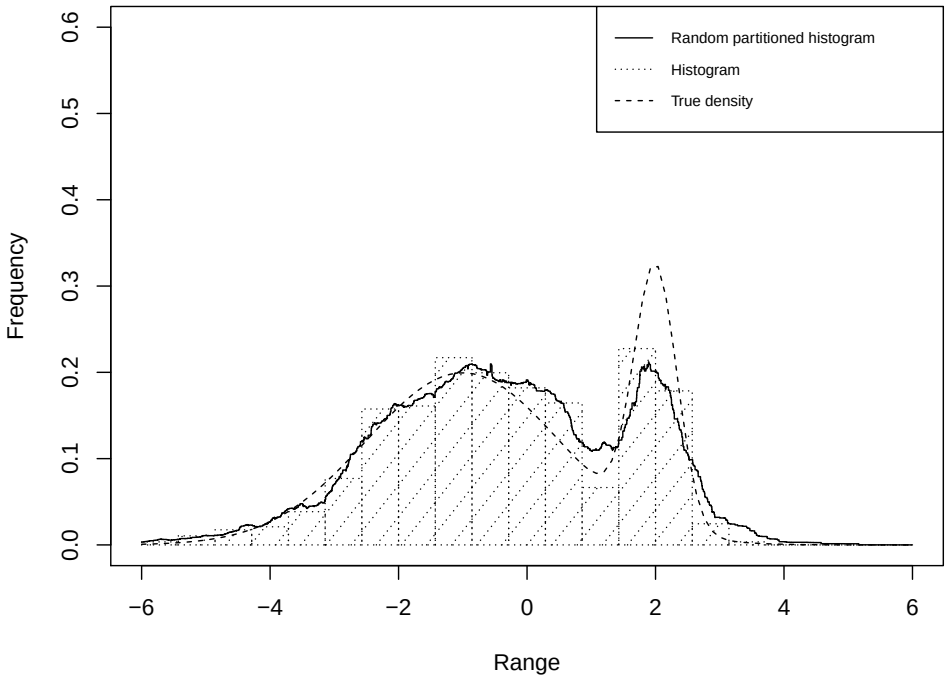

図 6.7　混合正規分布 $\frac{3}{4}N\left(-1,\ \left(\frac{3}{2}\right)^2\right)+\frac{1}{4}N\left(2,\ \left(\frac{1}{3}\right)^2\right)$, $(n=500)$ における RPH 推定量の例

まず、各データ数における Histogram と RPH の ISE 値と標準偏差の結果を示す。

表 6.8 は、データ数を変化させた時の ISE 値についての計算結果を示す。表で、比較して値が小さい方に下線を引いてある。Histogram と RPH のどちらも、n が増加するにつれて ISE 値は小さくなる。n に関わらず、RPH の方が ISE 値は小さい。

表 6.8　ISE 値 (多峰)

	$n=50$	$n=100$	$n=200$	$n=500$	$n=1000$	$n=5000$
Histogram	0.035698	0.028556	0.016088	0.012439	0.011044	0.003300
RPH $(r=30)$	<u>0.022530</u>	<u>0.017825</u>	<u>0.012525</u>	<u>0.008159</u>	<u>0.005257</u>	<u>0.001620</u>

表 6.9 は、データ数を変化させた時の ISE の標準偏差の計算結果を示す。表で、比較して値が小さい方に下線を引いてある。Histogram と RPH のどちらも、n が増加するにつれて ISE の標準偏差は小さくなる。$n=5000$ 以外では、RPH の方が ISE の標準偏差は小さい。$n=5000$ の場合には、Histogram のビン幅が十分小さく、ビン幅及びそれに伴うビン数がデータによって変動しにくいため、Histogram の方が RPH よりも分散が安定化すると考えられる。したがって、小、中標本においては、RPH の方が Histogram よりも分散は安定化している。

表 6.9　ISE 標準偏差 (多峰)

	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
Histogram	0.009934	0.005629	0.003052	0.003443	0.001543	<u>0.000188</u>
RPH ($r = 30$)	<u>0.004964</u>	<u>0.003822</u>	<u>0.002943</u>	<u>0.001964</u>	<u>0.001360</u>	0.000477

図 6.8 は各データ数で、ビン数を変化させた時の ISE 値である。n に関わらず、$m_{(RPH)}$ によって Histogram よりも推定精度が良いことが明らかになった。しかしながら、n が大きくなるにつれて改良の程度は小さくなっている。

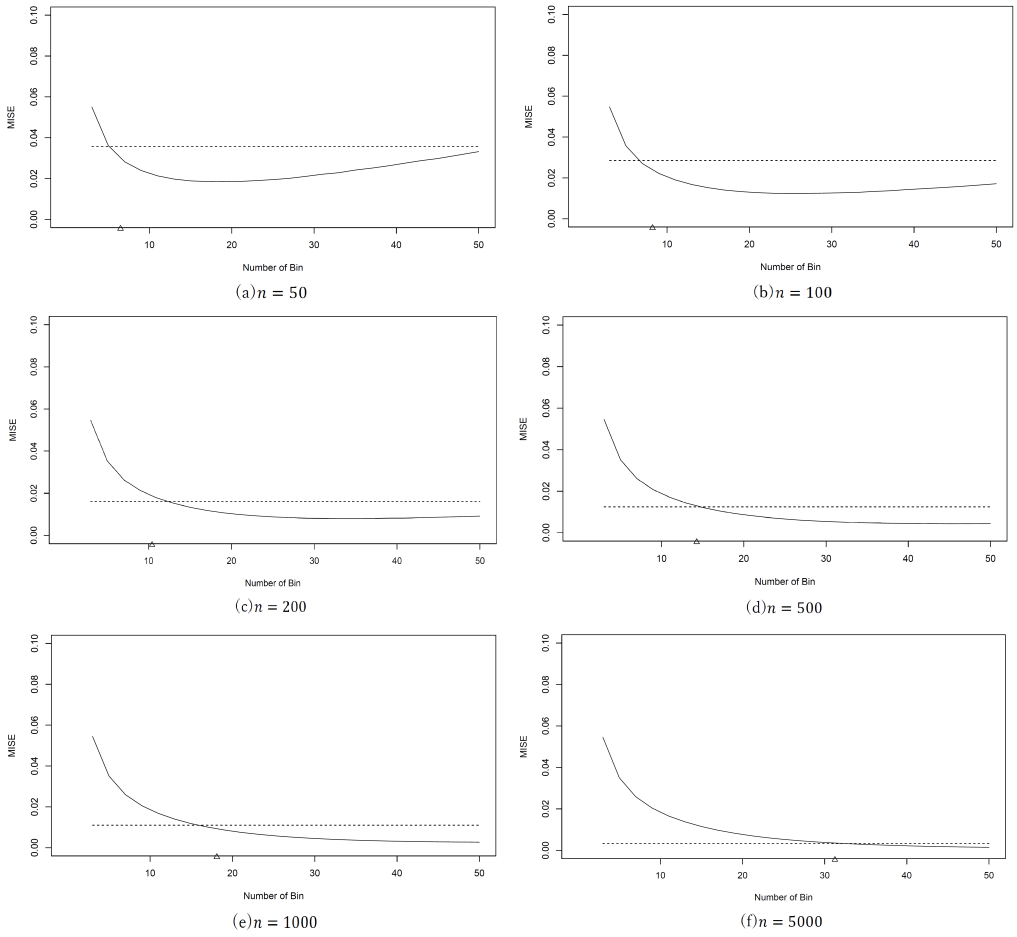

図 6.8　データ数ごとの各ビン数における ISE 値 (多峰、$r = 30$) (実線 : RPH、点線 : Histogram($m_{(RPH)} = Scott$)、三角印 : Histogram の平均ビン数)

表 6.10 は数値実験で得られたデータ数ごとの RPH の ISE 値の最小値及びその時のビン数、Histogram の数値実験における平均ビン数、RPH の ISE 値が最小時のビン数と Histogram の平均ビン数の比率である。Histogram と RPH どちらもビン数の推定にはスコットのルールを用いて

いる。n に関わらず、RPH の ISE 値が最小となる $m_{(RPH)}$ は mh より多いことが分かった。また、今回選択した n における $m_{(RPH)}/mh$ の平均は 3.07 であった。単峰の場合と同様に、多峰の場合も RPH では Histogram よりも多いビン数を選択する方が推定精度は良い傾向がある。

表 6.10　各データ数における ISE 値の最小値及びその時のビン数 (多峰、$r = 30$)

	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
最小 ISE 値	0.018446	0.012393	0.007955	0.004296	0.002649	0.000829
$m_{(RPH)}$	20	24	34	45	57	90
mh	6.48	8.23	10.43	14.28	18.10	31.24
$m_{(RPH)}/mh$	3.08	2.92	3.26	3.15	3.15	2.88

図 6.9 は数値実験で得られた単峰の分布における RPH の ISE 値が最小時のビン数と Histogram の平均ビン数の比 $m_{(RPH)}/mh$ (表 6.6)、および多峰の分布における数値実験で得られた $m_{(RPH)}/mh$ (表 6.10) それぞれに対する回帰直線である。横軸はデータ数で、$\log n$ で対数変換してある。グラフから、単峰の分布の場合、n が増加するにつれて $m_{(RPH)}/mh$ は 1 に近づいていく。一方で、多峰の分布の場合、n に関わらず $m_{(RPH)}/mh$ が約 3 である。この理由としては、Histogram のビン数の推定にスコットのルールを用いたことが挙げられる。スコットのルールでは参照分布で正規分布を仮定しており、単峰の分布に対しては当てはまりが良い。一方で、このルールから外れる多峰の分布に対しては平滑化過多のビン幅 (過小なビン数) を推定する傾向がある。したがって、分布の形状に関係なく RPH のビン数は Histogram よりも多い方が望ましく、単峰の場合には Histogram の 1.5 倍、多峰の場合には Histogram の 3 倍を目安にビン数選択することを推奨する。

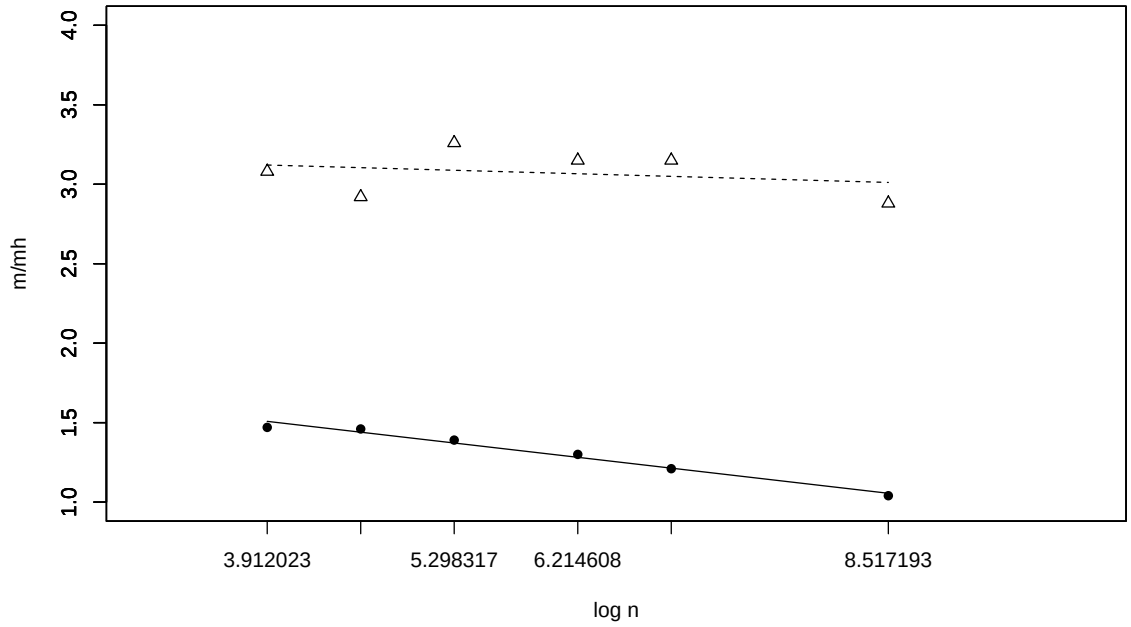

図 6.9　データ数ごとの $m_{(RPH)}/mh$ に対する回帰直線 (横軸：$\log n$、実線：$m_{(RPH)}/mh$(単峰) の回帰直線、破線：$m_{(RPH)}/mh$(多峰) の回帰直線、丸印：$m_{(RPH)}/mh$(単峰) の実現値、三角印：$m_{(RPH)}/mh$(多峰) の実現値)

7　結論と考察

7.1　結論

本稿では、特定の確率構造を仮定することなく、データに基づいて柔軟に適用可能な方法であるノンパラメトリックな確率密度推定法に関する一連の研究を行った。特に、ビン化によってデータを簡略化及び縮約化して推定するビン型の密度推定法である Histogram に注目した。ビン化によるデータの縮約化は非可逆的な変換であり、情報の損失を伴うため、その情報損失の回避を目的とした Histogram の改良を考えてきた。

3 章では、大規模データ解析に対する Histogram の改良として、1 次までの局所モーメント情報と各ビンの境界での 2 次連続性の条件を同時に満たす平滑化 Polynomial Histogram 推定量 (S-PH$_{(2,1)}$) を提案した。MISE 基準に基づく漸近的性質の導出から、その推定精度が $O\left(n^{-8/9}\right)$ であり、従来型の主なビン型推定法である Histogram の推定精度 $O\left(n^{-2/3}\right)$、1 次までの局所モーメント条件のみを満たす 1 次 PH の推定精度 $O\left(n^{-4/5}\right)$、2 次連続性の条件のみを満たす Histospline の推定精度 $O\left(n^{-6/7}\right)$、全データを用いた正規カーネルにおけるカーネル推定の精度 $O\left(n^{-4/5}\right)$ よりも理論的に優れた推定法であることを示した。また、S-PH$_{(2,1)}$ は漸近正規性が成り立つことを示した。有限標本特性を確かめるため、二山の分布を例に数値実験を行った。データ数が 10^3 以上の場合で、S-PH$_{(2,1)}$ が Histogram、1 次 PH、Histospline より優れた推定精度であることが明らかになった。各推定量の最適ビン数に注目すると、データ数が 10^5 の時、今回の実験における S-PH$_{(2,1)}$ のビン数は Histogram の約 1/45 である。

4 章では、2 次までの局所モーメント情報と各ビンの境界での 2 次連続性の条件を同時に満たす平滑化 Polynomial Histogram 推定量 (S-PH$_{(2,2)}$) を提案した。その漸近的性質の導出から、MISE の意味で推定精度が $O\left(n^{-10/11}\right)$ であり、Histogram、Histospline、2 次 PH の推定精度 $O\left(n^{-6/7}\right)$ よりも優れており、S-PH$_{(2,1)}$ を改良できることが示された。また、S-PH$_{(2,2)}$ は漸近正規性が成り立つことを示した。数値実験では、二山の分布を例に ISE の平均と標準偏差について従来型のビン型推定法と比較した。その結果、データ数が 10^4 以上の場合で、Hitogram、Histospline、2 次 PH、S-PH$_{(2,1)}$ より S-PH$_{(2,2)}$ の推定精度の方が優れていた。各推定量の最適ビン数に注目すると、データ数が 10^5 の時、今回の実験における S-PH$_{(2,2)}$ のビン数は Histogram の約 1/25 である。このことから、S-PH$_{(2,2)}$ は最も少ないビン数で、優れた推定精度を得られることが明らかになった。

5 章では、Histogram の推定区間に関わるビン幅の補正について議論した。既知の閉区間の定義域において、推定区間の始点を左端点に固定した最適ビン幅に基づく Histogram の推定区間が定義域と一致しない問題に着目した。理由としては、ビン幅は未知の母集団分布 f とデータ数 n によって決められる。したがって、ビン幅を決めた後に構成される推定区間は始点となる左端点の情報のみしか用いられず、結果として定義域の右端点と推定区間の右端点が異なる。そのため、この推定区間と定義域との間にずれが生じることを「ビン残差」と呼ぶことにする。この問題を解消す

る簡単な手順として、ビン残差を各ビンに等分配してビン幅を補正する手法を提案した。その漸近的性質について、補正後ビン幅で推定した Histogram が MISE の意味で漸近一致性と漸近正規性が成り立つことを示した。有限標本でのビン幅補正の特性を確かめるため、ISE の平均と標準偏差について数値実験を行った。右片側 tail 部分の確率が大きい場合、ビン残差部分にデータが多く存在するため、ビン幅補正による効果が特に大きく、データ数に関わらず、補正後ビン幅を用いた Histogram の MISE の方が、補正なし最適ビン幅を用いた Histogram より MISE と分散が共に小さくなり、推定量は改良される。

従来のビン幅推定で未知の母集団分布の推定とデータ数に依存する点について、小標本では推定量の精度の問題からビン幅推定量の安定した値を得られない。そのため、6 章では、データを繰り返し利用し、それと同時に母集団分布に依存しない (distribution-free な) ビン幅決定法を構築することで、小標本による不安定化と未知母集団分布の推定を同時に改良する方法を提案した。具体的には、分割点を一様乱数で決定した不等間隔なビン幅を用いた Histogram を繰り返し推定し、その平均を推定量とする Random Partitioned Histogram(RPH) を提案した。単峰及び多峰の分布を例に、データ数、ビン数、繰り返し回数に関して様々なパターンで ISE の平均と標準偏差について数値実験を行い、Histogram との比較から RPH の有限標本における性質を明らかにした。単峰と多峰のどちらの場合も、RPH に必要となるビン数、繰り返し回数を適切に調節することで、推定精度が改良され、分散も安定化することが明らかになった。

7.2 考察と今後の展開

(i) 小規模データに対する Histogram の改良

小規模データに対応するための Histogram の改良については、ビン幅補正法と Random Partitioned Histogram(RPH) の 2 つを議論した。

ビン幅の補正によって Histogram の推定区間と既知の閉区間の定義域を一致させることで、ビン幅を補正しない場合よりも推定精度を改良し、分散も安定化することが明らかになった。Histogram のビン幅補正法に関して理論面の整備は行ったため、今後は一般的によく利用されるオープンソースな R や Python 等の計算・統計処理プログラムソフトウェアでパッケージを実装していく。例えば、R の公式パッケージとして R Project へ登録し、広く一般において実データ分析に提案手法が適用されることを目指す。

ビン幅推定において必要となる未知の母集団分布に依存しない distribution-free な推定手法である RPH を提案し、数値実験から Histogram と比較して推定精度が改良できることを明らかにした。Histogram では最適ビン幅よりも狭いビン幅では平滑化不足となり、反対に、最適なビン幅より広いビン幅では平滑化過多で、どちらの場合も推定精度が悪くなる。RPH では、選択するビン数について広い許容範囲で Histogram より優れた推定精度を持つ。コンピュータプログラム上で計算が容易な推定法であり、ビン幅など統計の専門的知識に乏しい場合も含めて実用的に利用できる方法だと考えられる。

本稿では、有限標本の数値実験で Partitioned Histogram について示したが、ランダムな分割点

が理論的にはディリクレ分布に従うと考えられ、この理論面の整備については今後行う。RPH についてビン数や繰り返し回数を任意に設定したもとで議論したが、これらの最適な選択法についての理論面を明らかにする必要がある。また、既知の定義域を設定したが、一般に定義域は未知の場合が多い。RPH ではデータが少ない箇所で推定量が上方へ押し上げられる傾向が見られるが、定義域を広く取ることでこの傾向は軽減できると考えられる。今後は、RPH について理論面の整備を行い、理論的に最適なビン幅、繰り返し回数や定義域の広さを示し、それらの推定精度への影響や、推定法の有効性について明らかにする。

Histogram を改良することで、小規模データにおける情報不足による推定結果の不安定化を克服できることが本研究で示された。データ構造を見える化する単純で利用しやすい Histogram がモデルの根幹となるため、ビン幅補正法及び RPH の計算・統計処理プログラムへの実装は様々な分野に適用が期待できる。

(ii) 大規模データに対する Histogram の改良

大規模データに基づくビン型密度推定の観点から Histogram の研究を行ってきた。元データから簡単に計算でき、直観的にも利用しやすい局所モーメント情報の利用と平滑化を同時に満たすビン型推定量である S-PH は、従来型の主なビン型推定法よりも少ないビン数で優れた推定精度を得られることが明らかになった。また、ビン数が少ない、すなわち、各ビンにおいて広いビン幅で推定できるため、各ビンにおけるデータ数が多くなり、標本モーメント情報を安定的に利用できる推定法である。

	相対度数	局所平均	局所分散	分布関数の0次連続性	分布関数の1次連続性	分布関数の2次連続性
Histogram	○	×	×	○	×	×
1次PH	○	○	×	○	×	×
2次PH	○	○	○	○	×	×
Histospline	○	×	×	○	○	○
S-PH(2,1)	○	○	×	○	○	○
S-PH(2,2)	○	○	○	○	○	○

図 7.1　各推定量で満たす制約条件

図 7.1 は Histogram、PH、Histospline、S-PH の各推定量で満たす局所モーメントと連続性の条件についてまとめたものである。図中でその条件を満たすものは「○」、満たさないものは「×」で示してある。行で見ると、利用する付加情報が多く、ビンの境界での連続性が高いといった「○」の数が多い推定量の方が推定精度が優れていることが分かる。また、S-PH$_{(2,1)}$ と S-PH$_{(2,2)}$ の理論研究から、モーメントの制約と連続性の制約条件は方程式として従属関係がなく、異なる情報構造を持つと判断できる。すなわち、寒河江 (2022) で指摘しているようにモーメントの次数と連続

性の次数をそれぞれ上げることで同時に推定精度の改良が可能であることが本研究で確認できた。この考察を更に展開すると、精度保障に必要なオーダーに対応したモーメント次数あるいは連続性の次数を決めて、それに対応させた S-PH$_{(p,q)}$ について、p 次の連続性とデータから計算する q 次までのモーメント情報によって精度を前もって保障できることになる。

本稿では、S-PH の構築時に、利用する局所モーメント情報について 1 次までと 2 次までの 2 通りの場合を扱った。そのため、2 次連続性と高次 (q 次) までの局所モーメント情報を同時に満たす推定モデル S-PH$_{(2,q)}$ の表現とその理論的特性について今後明らかにする。また、各ビンの境界における連続性について、一般的に平滑化としてよく用いられる 2 次連続性に固定して議論した。p 次連続性と q 次局所モーメント情報を同時に高次へ拡張した S-PH$_{(p,q)}$ の一般化表現について理論面の整備を今後検討する。S-PH では、推定量の非負性を緩和することで推定精度を改良する方法であるため、非負性は担保していない。非負性を保つような S-PH についても今後検討を進めていく。

大規模データの分析に対しては、ハードウェアとソフトウェアの組み合わせによる工夫と、計算アルゴリズムを工夫する 2 つのアプローチが主に挙げられるが、3 つ目のアプローチとして統計手法の工夫がある。その一つの例が、小暮・寒河江 (2010) が提案したカーネル・データスカッシングである。DuMouchel(2002) によるデータスカッシングとは、元の大規模データと近似的に同一の分析結果、すなわち、元の大規模データとモーメントが一致するように代替的な小規模データを構成することで、分析時の計算負荷を緩和する方法である。局所モーメント情報を利用し、各ビンの推定から構成される S-PH は、このデータスカッシングと似た発想であり、ビン化によってデータを縮約して計算量を軽減しつつ、元データが持つ情報を保持できる方法だと考えられる。一般的にデータ量が数十テラバイトから数ペタバイトの範囲に及ぶと言われているビッグデータの解析においては、S-PH の少ないビン数で優れた推定精度を達成できる性質が有効に働くと思われる。

参考文献

[1] Boneva, L.I., Kendall, D. and Stefanov, I. (1971), "Spline Transformations: Three New Diagnostic Aids for the Statistical Data-Analyst." *Journal of the Royal Statistical Society. Series B (Methodological)*, 33.1, pp.1-71.

[2] Bowman, A.W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates", *Biometrika*, Vol.71, pp.33-36.

[3] DiCiccio, T.J. and Efron, B. (1996), "Bootstrap confidence intervals", *Statistical science*, 11(3), pp.189-228.

[4] Doane, D.P. (1976), "Aesthetic Frequency Classifications", *The American Statistician*, Vol.30, No.4, pp.181-183.

[5] DuMouchel, W. (2002), "Data Squashing: Constructing summary data sets", in *Handbook of Massive Data Sets*, Springer, pp.579-591.

[6] Freedman, D. and Diaconis, P. (1981), "On The Histogram as a Density Estimator: L_2 Theory", *Zeitschrift fuer wahrsheinlickkeitstheorie und Verwandle Gebeite*, Vol.57, pp.453-476.

[7] Jones, M.C., Samiuddin, M., AL-Harbey, A.H. and Maatouk, T.A.H. (1998), "The edge frequency polygon", *Biometrika*, Vol.85, Issue 1, pp.235–239.

[8] Kogure, A. (1987), "Asymptotically Optimal Cells for a Histogram", *The Annals of Statistics*, Vol.15, No.3, pp.1023-1030.

[9] Lecoutre, Jean-Pierre (1987), "The Histogram with Random Partition", in *New Perspective in Theoretical and Applied Statistics*(Editors: M. L. Puri, J. P. Vilaplana and W. Wertz), John Wiley & Sons, pp.265-276.

[10] Lii, Keh-Shin and Rosenblatt, M. (1975), "Asymptotic behavior of a spline estimate of a density function", *Computers & Mathematics with Applications*, 1(2), pp.223-235.

[11] Minnotte, M.C. (1996), "The Bias-Optimized Frequency Polygon", *Computational Statistics*, 11, pp.35-48.

[12] Minnotte, M.C. (1998), "Achieving Higher-Order Convergence Rates for Density Estimation with Binned Data", *Journal of the American Statistical Association*, Vol.93, No.442, pp.663-672.

[13] Rudemo, M. (1982), "Empirical Choice of Histogram and Kernel Density Estimatoras", *Scandinavian Journal of Statistics*, Vol.9, pp.65-78.

[14] Sagae, M. and Scott, D.W. (1997), "Bin Interval Method of Locally Adaptive Nonparametric Density Estimation", Joint Statistical Meetings 1997.

[15] Schoenberg, I.J. (1973), "Splines and Histograms", *Spline Functions and Approximation Theory*, Birkhauser, Basel, pp.277-327.

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

[17] Scott, D.W. (1985), “Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions”，*The Annals of Statistics*，Vol.13，No.3，pp.1024-1040.

[18] Scott, D.W. (1985), “Frequency Polygons: Theory and Application”, *Journal of the American Statistical Association*, Vol. 80, No. 390, 348-354.

[19] Scott, D.W. and Terrell, G.R. (1987), “Biased and Unbiased Cross-Validation in Density Estimation”, *Journal of the American Statistical Association*, Vol.82, pp.1131-1146.

[20] Scott, D.W. and Sagae, M. (1997), “Adaptive Density Estimation with Massive Data Sets”, Joint Statistical Meetings 1997.

[21] Sibuya, M. and Yamato, H. (1995), “Characterization of Some Random Partitions”，*Japan Journal of Industrial and Applied Mathematics*，12，pp.237-263.

[22] Sturges, H.A. (1926), “The Choice of a Class Interval”, *Journal of the American Statistical Association*, Vol.21, pp.65-66.

[23] Terrell, G.R. and Scott, D.W. (1985), “Oversmoothed nonparametric density estimates”, *Juornal of the American Statistical Association*, Vol.80, pp.209-214.

[24] Terrell, G.R. and Scott, D.W. (1992), “Variable Kernel Density Estimation”，*The Annals of Statistics*，Vol.20，No.3，pp.1236-1265.

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

[26] Wand, M.P. (1997), “Data-Based Choice of Histogram Bin Width”, *The American Statistician*, Vol.51, No.1, pp.59-64.

[27] 小暮, 寒河江 (2010), “カーネル・データスカッシング—カーネル密度推定法のデータマイニングへの応用—”, 日本統計学会誌, 第 39 巻, 第 2 号, pp.243-263.

[28] 齊藤, 寒河江 (2021), “3 次スプライン関数によるヒストグラム平滑化とその漸近的性質～Boneva, Kendall and Stefanov 型と Lii and Rosenblatt 型モデルの理論的同等性～”, 人間社会環境研究, 第 41 号, pp.49-62.

[29] 齊藤, 寒河江 (2021), “ヒストグラムのランダムな分割に基づくビン幅決定法のシミュレーション研究”, 人間社会環境研究, 第 42 号, pp.213-226.

[30] 齊藤, 寒河江 (2022), “多項式型 Histogram の平滑化とその漸近的性質”, 金沢大学人間社会研究域ディスカッションペーパー, No.68.

[31] 齊藤, 寒河江 (2022), “Spline Interpolation of Polynomial Histogram Density Function and Its Asymptotic Properties”, 金沢大学人間社会研究域ディスカッションペーパー, No.69.

[32] 齊藤, 寒河江 (2022), “有限区間における Histogram のビン幅補正について”, 金沢大学人間社会研究域ディスカッションペーパー, No.70.

[33] 寒河江, 齊藤, 原田, 今井 (2019), “いしかわ・金沢 風と緑の楽都音楽祭 2019 実態分析報告”,

金沢大学.

[34] 寒河江 (2022), “データ集約型のノンパラメトリック密度推定法について”, 2022 年度 統計関連学会連合大会.