統計数理(2005)
第 53 巻 第 2 号 375–389

特集「計算推論—モデリング・数理・アルゴリズム—」
[原著論文]

パーセント点に集計されたデータからの密度関数の推定

—バイアス・パズルの考察—

小暮 厚之[1] ・寒河江 雅彦[2]

(受付 2005 年 3 月 1 日；改訂 2005 年 6 月 30 日)

要　　旨

統計資料が公開される際に，原データの取り得る範囲をいくつかの区間に分割し，各区間ごとの度数データにまとめられることが多い．代表的な分割法に，各区間の長さが等しくなるように分割する方法と各区間の度数が等しくなるように分割する方法とがある．後者の分割法は，データを十分位数のようなパーセント点に集計することであり，データが密な領域では区間幅が狭く，疎な領域では区間幅が広くなるという特性を持つ．しかし，密度関数推定という観点から考えるとき，この等度数分割による度数表示が著しく大きなバイアスを引き起こす可能性についてはあまり知られていないようである．本稿では，このバイアス問題を理論的に考察し，そのひとつの対処法として，パーセント点に基づくビン化カーネル密度推定法を提案する．提案した手法の漸近的な性質を導くとともに，そのパフォーマンスをシミュレーションによって例示する．

キーワード：パーセント点，度数データ，ヒストグラム，カーネル推定量，バイアス.

1. はじめに

統計資料が公開される際に，原データの取り得る範囲をいくつかの区間に分割し，各区間ごとの度数データに集計することが多い．代表的な分割法に，各区間の長さが等しくなるように分割する方法(等区間幅分割)と各区間の度数が等しくなるように分割する方法(等度数分割)とがある．後者の分割法は，データを十分位数のような等パーセント点に集計することに他ならない．それは，データが密な領域では区間幅が狭く，疎な領域では区間幅が広くなる特性を持つ．例えば，総務省による平成 15 年度家計調査では，調査した 4464 勤労世帯の年間収入を表 1 の等間隔に分割した区間の度数データとともに表 2 にある十分位数という形で公開している．ただし，表 1 は表 2 の十分位数による分割との対比のために，オリジナルな表より粗く区分してある．ここで，表 2 の各区間の度数が等しくない理由は，観測値を記録する際の丸め誤差によって，同一の値を取るデータが複数存在するためであろう．

この 2 つの表をグラフ表示したものが図 1 と図 2 である．図 1 は通常のヒストグラムである．ヒストグラムの形状は，区間に集計される前のオリジナルなデータの確率分布の特徴を反映し

[1] 慶應義塾大学 総合政策学部：〒252–8520 神奈川県藤沢市遠藤 5322
[2] 岐阜大学 工学部：〒501–1193 岐阜市柳戸 1–1

表 1. 勤労者世帯収入データの階級集計.

階級下限	階級上限	度数
0	200	40
200	300	204
300	400	428
400	500	610
500	600	629
600	700	592
700	800	523
800	900	407
900	1000	296
1000	1250	444
1250	1500	168
1500	—	124

表 2. 勤労者世帯収入データのパーセント点集計.

階級下限	階級上限	度数
0	362	499
362	445	455
445	514	442
514	586	441
586	657	437
657	738	450
738	825	427
825	950	448
950	1134	422
1134	—	443

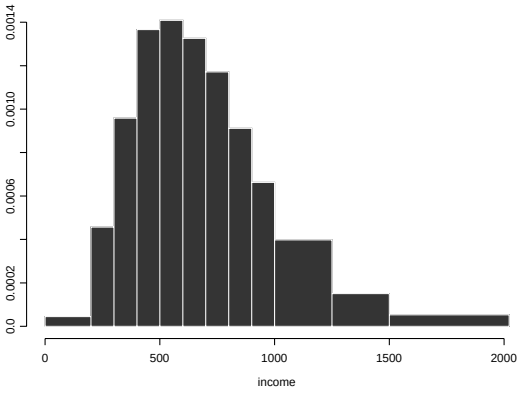

図 1. 通常のヒストグラム：勤労家計収入データ.

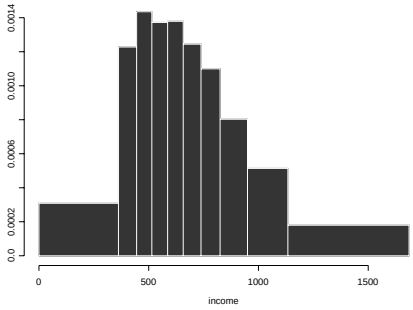

図 2. パーセント点ヒストグラム：勤労家計収入データ.

ている．実際，階級幅が十分小さければ，ヒストグラムはオリジナルなデータの確率密度関数の推定値となる．図 2 は十分位数に基づくパーセント点ヒストグラムである．図 1 とはやや異なる印象を与える．

通常のヒストグラムと同様に，パーセント点に基づくヒストグラムも，オリジナルな分布の密度関数の推定値と考えることができる．通常のヒストグラムが，各階級の区間幅を一定にして相対度数を表示しているのに対して，パーセント点ヒストグラムは，各階級の度数を等しくするように区間幅を調整して相対度数を表示する．このため，一般に，パーセント点ヒストグラムは，データの疎な領域(例えば，分布の両裾)でより広いビン幅を与え，データの密な領域(例えば，分布の中心)でより狭いビン幅を与える．このパーセント点ヒストグラムは，局所的に階級区間幅を変化させていると言う点で，いわゆる適応型ヒストグラム(adaptive histogram)の一種と考えられる．Kogure (1987)や Terrell and Scott (1992)で議論されているように，適切に区間幅を調整することによって，等間隔区間のヒストグラムより適応型ヒストグラム推定効率の方が高くなる．しかし，Lecoutre (1987)や Scott (1992)の考察から，パーセント点ヒス

トグラムを用いると，正規分布を含む多くの分布に対して，バイアスが大きくなるという問題点が示唆される．本稿では，このバイアス問題を取り上げ，その問題点を理論的に考察するとともに，カーネル法による改良を提案する．

2. ヒストグラムのバイアス解析

2.1 分割法と 2 種類のヒストグラム

$\{X_1, X_2, \ldots, X_n\}$ を確率分布 F からの大きさ n の無作為標本とする．データを，基準点が x_0，幅が δ の階級区間に集計する場合は，データは各区間

$$\nu_j \equiv [x_0 + j\delta - (\delta/2),\, x_0 + j\delta + (\delta/2)), \quad j = 0, \pm 1, \pm 2, \ldots$$

に落ちる度数

$$N_j \equiv \sum_{i=1}^{n} I\left(X_i \in \nu_j\right), \quad j = 0, \pm 1, \pm 2, \ldots \tag{2.1}$$

に縮約される．ここで，$I(A)$ は A が真であるときに 1，偽であるときに 0 を取る指示関数とする．階級区間 $\{\nu_j\}$ に対するヒストグラムは

$$H(x|\delta) \equiv \frac{N_j}{n\delta} = \frac{F_n\left(x_0 + j\delta + (\delta/2)\right) - F_n\left(x_0 + j\delta - (\delta/2)\right)}{\delta}, \quad x \in \nu_j \tag{2.2}$$

と定義される．ここで，$F_n(\cdot)$ は $\{X_i\}$ の経験分布関数

$$F_n(x) \equiv \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x), \quad -\infty < x < \infty$$

である．これは階級幅が一定である通常のヒストグラムである．

一方，データをパーセント点にまとめる集計法では，区間幅を固定する代わりに，各ビンに入る観測値の度数を一定に固定しようとする．$\{X_i\}$ を m 等分する場合，データはパーセント点

$$Y_j \equiv F_n^{-1}(j/m) \equiv \inf\{x : F_n(x) \ge j/m\}, \quad j = 1, 2, \ldots, m \tag{2.3}$$

に縮約される．さらに

$$Y_0 \equiv \min_{1 \le i \le n} \{X_i\}, \quad Y_m \equiv \max_{1 \le i \le n} \{X_i\}$$

とすると，対応する階級区間は

$$\Pi_j \equiv [Y_{j-1}, Y_j), \quad j = 1, 2, \ldots, m-1;\ \Pi_m \equiv [Y_{m-1}, Y_m]$$

となる． Π_j に含まれる観測値

$$M_j \equiv \sum_{i=1}^{n} I\left(X_i \in \Pi_j\right) \tag{2.4}$$

は，タイ・データがない場合には，$[n/m]$ または $[n/m] + 1$ 個の観測値を含む．ここで，$[\cdot]$ はガウス記号である．パーセント点階級区間 $\{\Pi_j\}$ に対するヒストグラムは

$$H_{\%}(x|m) \equiv \frac{M_j}{n(Y_j - Y_{j-1})}, \quad x \in \Pi_j \tag{2.5}$$

である．これをパーセント点ヒストグラムと呼ぶ．これは，通常のヒストグラムと異なり，階級幅が変化する．

2.2 MISE 基準

2 種類のヒストグラム(2.2)及び(2.5)を密度関数 $f(x) \equiv dF(x)/dx$ の推定量と考えるとき，その効率性を測る代表的な基準は，平均平方誤差の積分(MISE，mean integrated squared error)

$$\mathrm{MISE} \equiv \int_{-\infty}^{\infty} \mathrm{E}[(\widehat{f}(x) - f(x))^2] dx$$

である(密度関数推定における MISE 基準の役割と意味については，Jones (1991)を参照されたい)．ここで，$\widehat{f}$ は(2.2)または(2.5)である．MISE は 2 乗バイアスの積分(ISB，integrated squared bias)

$$\mathrm{ISB} \equiv \int_{-\infty}^{\infty} (\mathrm{E}[\widehat{f}(x)] - f(x))^2 dx$$

と分散の積分(IV，integrated variance)

$$\mathrm{IV} = \int_{-\infty}^{\infty} \mathrm{Var}(\widehat{f}(x)) dx$$

の和として

$$\mathrm{MISE} = \mathrm{ISB} + \mathrm{IV}$$

と表せる．

2.3 バイアス

一定の条件の下で，通常のヒストグラムの期待値は，$\delta \to 0$ のとき

$$\mathrm{E}[H(x|\delta)] \approx f(x) + \left((j-1)\delta + \frac{\delta}{2} - x \right) + O(\delta^2), \quad x \in \nu_j, \quad \delta \to 0$$

と計算され，その ISB の漸近的表現は

$$\mathrm{ISB} \equiv \int_{-\infty}^{\infty} (\mathrm{E}[H(x|\delta) - f(x)])^2 \, dx \sim \frac{\delta^2}{12} R(f') \tag{2.6}$$

と与えられる(例えば，Scott, 1992 を見られたい)．ここで，$R(\cdot)$ は任意の関数 g に対して，その L^2 距離

$$R(g) \equiv \int_{-\infty}^{\infty} g(x)^2 dx$$

を表す．Freedman and Diaconis (1981)によって示されているように，この結果が成立するために本質的に重要な仮定は，$R(f) < \infty$ が成立することである．

一方，パーセント点ヒストグラムの期待値は，$m \to \infty$ のとき

$$\mathrm{E}[H_{\%}(x|m)] \approx f\left(F^{-1}\left(\frac{j - 1/2}{m} \right) \right) + O(m^{-2}), \quad x \in \Pi_j, \quad m \to \infty$$

と計算される．Lecoutre (1987)は，

$$R(f'/f) \equiv \int_{-\infty}^{\infty} \left(\frac{f'(x)}{f(x)} \right)^2 dx < \infty$$

という条件の下で，パーセント点ヒストグラムの ISB が

$$\mathrm{ISB} \equiv \int_{-\infty}^{\infty} (\mathrm{E}[H_{\%}(x|m) - f(x)])^2 \, dx \sim \frac{1}{12m^2} R\left(\frac{f'}{f} \right) \qquad (m \to \infty) \tag{2.7}$$

と漸近的に表現できることを示している．

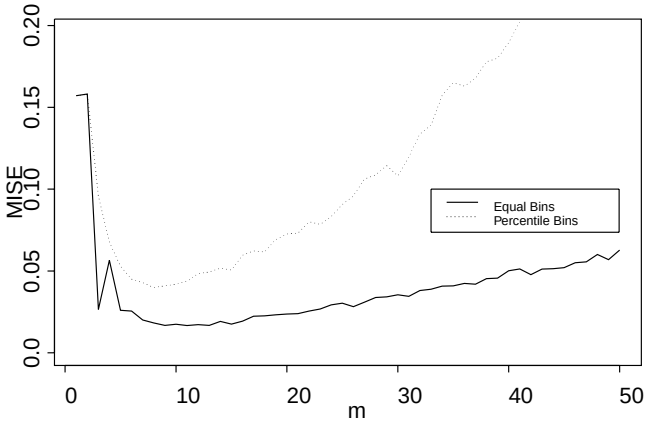

図 3. $n = 100$ **に対する** MISE **の比較：パーセント点ヒストグラム（点線）**VS. **通常のヒストグラム（連続線）．**

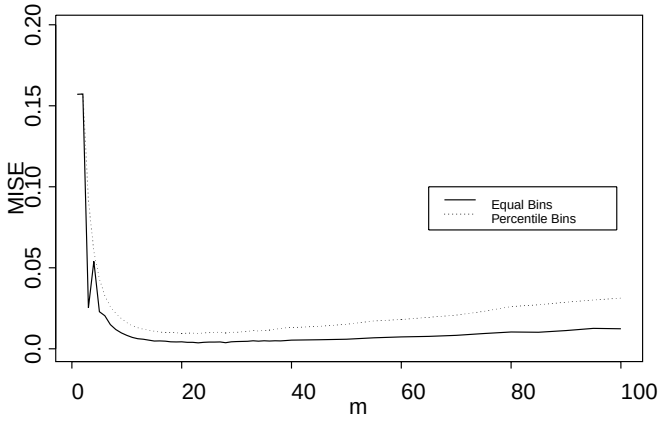

図 4. $n = 1000$ **に対する** MISE **の比較：パーセント点ヒストグラム（点線）**VS. **通常のヒストグラム（連続線）．**

Scott (1992, p. 70)**は，適応型ヒストグラムの議論の中で，**(2.7)**と同一の表現を導き，** f **が標準正規分布ならば，**

$$R(f'/f) = \int_{-\infty}^{\infty} x^2 dx = \infty \quad !$$

となることを指摘し，“puzzling”**という言葉で表現している．この「バイアス・パズル」は，データが正規分布に従うとき，有限サンプルにおいても，パーセント点ヒストグラムが著しく大きなバイアスを持つことを示唆する．**Scott **は，シミュレーションにより有限サンプルに対する**

MISE を計算し，通常のヒストグラムに比べて，パーセント点ヒストグラムのパフォーマンスが劣ることを例示している．

図 3 と図 4 は，この Scott の例に倣って，改めて計算した結果である．$[-4,4]$ 上で切断した標準正規分布から，大きさ $n=100$ 及び $n=1000$ の無作為標本を生成し，平方誤差積分

$$\mathrm{ISE} \equiv \int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx$$

を計算した．このような計算を繰り返し，その算術平均を求めれば，MISE を推定できる．この計算を 100 回繰り返したときのパーセント点ヒストグラム（点線）と通常のヒストグラム（通常の連続線）の ISE の算術平均値を表している．

ここで，パーセント点ヒストグラムは標本パーセント点（m 分位数）に基づいている．また，通常のヒストグラムは，階級区間の個数をパーセント点ヒストグラムの個数 m と同一とするように，区間幅を $\delta = 8/m$ と設定した．Scott の計算結果と同様に，m の全範囲にわたってパーセント点ヒストグラムの MISE が通常のヒストグラムより大きい値を取っていることが分かる．

2.4 分散

通常のヒストグラムの分散は

$$\mathrm{Var}(H(x|\delta)) \approx \frac{f(x)}{n\delta}, \quad x \in \nu_j$$

と計算され，その IV は

$$\mathrm{IV} \equiv \int \mathrm{Var}\left(H(x|\delta)\right) dx \sim \frac{1}{n\delta} \int f(x) dx = \frac{1}{n\delta} \tag{2.8}$$

と与えられる（Scott, 1992）．一方，パーセント点ヒストグラムの分散は

$$\mathrm{Var}(H_{\%}(x|m)) \approx \frac{m}{n} \times f\left(F^{-1}\left(\frac{j-1/2}{m}\right)\right)^2, \quad x \in \Pi_j$$

と計算され，その IV は

$$\mathrm{IV} \equiv \int_{-\infty}^{\infty} \mathrm{Var}\left(H_{\%}(x|m)\right) dx \sim \frac{m}{n} R(f) \tag{2.9}$$

と与えられる（Lecoutre, 1987）．正規分布を含む多くの代表的な分布について $R(f) < \infty$ が成立するため，パーセント点ヒストグラムの IV に関しては，ISB のような深刻な問題は生じない．

2.5 MISE

(2.6) と (2.8) より，通常のヒストグラム (2.2) の MISE の漸近的表現は，

$$\mathrm{MISE} = \mathrm{ISB} + \mathrm{IV} \sim \frac{\delta^2}{12} R(f') + \frac{1}{n\delta}$$

と与えられる．MISE を最小にするような δ のオーダーは $n^{-1/3}$ であり，対応する MISE のオーダーは $n^{-2/3}$ となる．すなわち，MISE 基準で測定した通常のヒストグラムの推定効率性は $n^{-2/3}$ のオーダーである．

一方，パーセント点ヒストグラムの MISE は，(2.7) と (2.9) より

$$\mathrm{MISE} = \mathrm{ISB} + \mathrm{IV} \sim \frac{1}{12m^2} R(f'/f) + \frac{m}{n} R(f)$$

と与えられる．MISE を最小にするような m のオーダーは $n^{1/3}$ であり，対応する MISE のオーダーは，通常のヒストグラムと同じく，$n^{-2/3}$ となる．

3. バイアス・パズルの考察

例えば，対数正規分布が当てはまるようなデータでは，通常の等区間幅の度数表示では明らかに不十分であろう．そのような場合に，等度数分割は，データの密な部分は詳細に，疎な部分は大まかに記述できる直感的な手順としてよく使われている．このように，前節で指摘したバイアス・パズルを考慮すべき多くの統計資料が存在する．

Lecoutre (1987)は，$R(f'/f)<\infty$ という仮定の下で，パーセント点ヒストグラムの ISB の漸近的表現(2.7)を与えている．この結果を逆に考えると．正規分布のように $R(f'/f)=\infty$ となる分布に対しては，ISB の漸近表現は(2.7)のように必ずしも表されないことが示唆される．この点を一般的に議論する代わりに，$R(f'/f)=\infty$ となる 2 つの具体的な確率分布を取り上げ，その ISB を計算する．考察を簡潔にするために，(2.5)において，標本パーセント点(2.3)の代わりに，母集団パーセント点

$$y_j \equiv F^{-1}(j/m), \quad j=1,2,\ldots,m-1 \tag{3.1}$$

を用いたパーセント点ヒストグラムを考える．ただし，y_0 及び y_m は分布 F の端点とする．

例 1. 密度関数

$$f(x)=2x, \quad 0\le x\le 1 \tag{3.2}$$

を考える．その $R(f'/f)$ は

$$R(f'/f)=\int_0^1 \frac{1}{x^2}dx=\infty$$

であり，逆分布関数は $F^{-1}(z)=\sqrt{z}, \quad 0\le z\le 1$ と与えられる．従って

$$\mathrm{E}[H_{\%}(x|m)]=\frac{F(y_j)-F(y_{j-1})}{y_j-y_{j-1}}=\frac{1/m}{\sqrt{j/m}-\sqrt{(j-1)/m}}, \quad x\in\Pi_j$$

となる．但し，$y_0=0$，$y_m=1$ とする．その結果，(3.2)の ISB は，

$$\mathrm{ISB}=\int_0^1 (\mathrm{E}[H_n(x)]-f(x))^2\,dx=\gamma m^{-3/2}+O(m^{-2}),$$

と計算できる．ここで，γ は定数

$$\gamma=\frac{133}{320}-\frac{5}{128}\int_1^\infty b_4(x)x^{-7/2}dx,$$

であり，$b_4(x)$ は 4 次ベルヌイ関数である．証明は，付録 A を見られたい．従って，$R(f'/f)=\infty$ であるにもかかわらず，ISB は有限な値を取る．ただし，そのゼロへの収束オーダーは $m^{-3/2}$ であり，(2.7)の漸近表現のオーダーよりも遅い．このとき，MISE を最小にする m のオーダーは $n^{2/5}$ となり，対応する MISE の収束オーダーは $n^{-3/5}$ となる．その結果，通常のヒストグラムに比べて，漸近的な推定効率が著しく低下する．

例 2. 平均が 1 の指数分布の密度関数

$$f(x)=e^{-x}, \quad x>0 \tag{3.3}$$

を考える．この密度関数に対して

$$R(f'/f)=\int_0^\infty 1dx=\infty$$

であり，逆分布関数は $F^{-1}(z) = -\log(1-z), \quad 0 \le z < 1$ である．このとき

$$\mathrm{E}[H_n(x)] = \frac{1/m}{\log(1-(j-1)/m) - \log(1-j/m)} = \frac{1}{m\log(1+1/(m-j))}$$

であり

$$\begin{aligned}\mathrm{ISB} &= \int_0^\infty (\mathrm{E}[H_n(x)] - f(x))^2\, dx \\ &= \frac{1}{m^2}\sum_{j=1}^{m}\left[(m-j) + 1/2 - \frac{1}{\log(1+1/(m-j))}\right]\end{aligned}$$

と計算される．ここで，$\log(1+x) \le x$ に注意すれば

$$\mathrm{ISB} \le \frac{1}{m^2}\sum_{j=1}^{m}\left[(m-j) + 1/2 - \frac{1}{1/(m-j)}\right] = \frac{1}{2m},$$

を得る．従って，例 1 と同じく，$R(f'/f) = \infty$ であっても，ISB は有限の値にとどまる．これは，たとえ $R(f'/f)$ が無限になろうともパーセント点ヒストグラムの ISB がゼロに収束していく別の例である．

以上の 2 つの例は，$R(f'/f) = \infty$ の場合には，ISB がゼロに収束するスピードが，通常期待されるオーダーである m^{-2} よりも遅いことを示唆する．言い換えると，Scott (1992)が指摘したバイアス・パズルとは，ISB が無限大になることではなく，その収束オーダーが通常想定される収束オーダーよりも遅いことであると理解できる．

4. パーセント点に基づくカーネル推定

3 節の考察は，パーセント点に基づいて確率密度を推定する場合には，$R(f'/f) < \infty$ でない限り，一般に大きなバイアスが予想されることを示唆する．正規分布を含む多くの分布に対して $R(f'/f) < \infty$ は成立しないため，このことは，パーセント点に集計されたデータにヒストグラム法を適用することの危険性を指摘する．本節では，カーネル法を用いることによって，バイアス問題が緩和されることを議論する．

4.1 カーネル推定量

もしも標本全体 $\{X_1, X_2, \ldots, X_n\}$ が利用可能ならば，通常のカーネル推定量

$$\widehat{f}(x|h) \equiv \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \quad -\infty < x < \infty, \tag{4.1}$$

を用いることができる．ここで，$K_h(\cdot)$ は $K_h(u) \equiv K(u/h)/h, \quad -\infty < u < \infty$ と定義される．カーネル関数 K は，原点を中心として左右対称な有界関数であり

$$\int_{-\infty}^{\infty} K(u)du = 1, \quad \int_{-\infty}^{\infty} |K(u)|du < \infty$$

を満たすものとする．また，$h = h_n$ は，バンド幅と呼ばれる

$$\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} nh_n = \infty$$

となる非確率的正数列 $\{h_n\}$ を表す．実際には，カーネル関数 K として原点を中心として左右対称な密度関数を用いた 2 次カーネル推定量を用いることが多い．その代表例は K を標準正規密度関数とした正規カーネル推定量である．

データが予め階級区間 $\{\nu_j\}$ に集計されているとき，いわゆるビン化カーネル推定量(binned kernel estimator)

$$\widetilde{f}(x|h,\delta) \equiv \frac{1}{n}\sum_{j=-\infty}^{\infty} N_j K_h(x-g_j), \quad -\infty < x < \infty, \tag{4.2}$$

を計算することができる．ここで，N_j は(2.1)で定義される．また，グリッド点 $\{g_j\}$ は，区間 ν_j の中点

$$g_j \equiv x_0 + j\delta, \quad j = 0, \pm 1, \pm 2, \ldots$$

である．このような，ビン化カーネル推定量については，Fan and Marron (1994)及び Hall and Wand (1996)を見られたい．

データがパーセント点に集計されている場合には，ビン化カーネル推定量に倣って

$$\widetilde{f}_{\%}(x|h,m) \equiv \frac{1}{n}\sum_{j=1}^{m-1} D_j K_h(x-Y_j), \quad -\infty < x < \infty, \tag{4.3}$$

という推定量を考えることができる．ここで，$\{Y_j\}$ は(2.3)の標本パーセント点であり，$\{D_j\}$ は

$$D_j \equiv \frac{M_j + M_{j+1}}{2}, \quad j = 1, 2, \ldots, m-1$$

と定義される．また，M_j は(2.4)で定義される．これをパーセント点カーネル推定量と呼ぶ．パーセント点カーネル推定量は，通常のビン化カーネル推定量と異なり，グリッド点 $\{Y_j\}$ は一様に配置されていない．

以下本節では，パーセント点カーネル推定量がバイアス問題を改良するか否かを理論的に考察する．このため，3 節と同様に，母集団パーセント点(3.1)が既知であり，グリッド点は $\{y_j\}$，M_j は区間 $[y_{j-1}, y_j)$ の度数とする．

4.2 バイアス

パーセント点カーネル推定量(4.3)の期待値は

$$\mathrm{E}[\widetilde{f}_{\%}(x|h,m)] = \frac{1}{n}\sum_{j=1}^{m-1} \mathrm{E}[D_j] K_h(x-y_j) = \frac{1}{m}\sum_{j=1}^{m-1} K_h(x-y_j)$$

と計算される．通常のカーネル推定量(4.1)の期待値は

$$\mathrm{E}[\widehat{f}(x|h)] = \int_{-\infty}^{\infty} K_h(x-y) f(y) dy = \int_0^1 K_h\left(x - F^{-1}(z)\right) dz$$

と表せる．ここで，$c_j \equiv j/m - 1/2m$ $(j = 1, 2, \ldots, m)$ とすると

$$\frac{1}{m}\sum_{j=1}^{m-1} K_h(x-y_j) = \sum_{j=1}^{m-1} \int_{c_j}^{c_{j+1}} K_h\left(x - F^{-1}(j/m)\right) dz$$

と表現できることに注意すると：

$$\begin{aligned} &|\mathrm{E}[\widetilde{f}_{\%}(x|h,m)] - \mathrm{E}[\widehat{f}(x|h)]| \\ &= \left| \sum_{j=1}^{m-1} \int_{c_j}^{c_{j+1}} K_h(x - F^{-1}(j/m)) - K_h(x - F^{-1}(z)) dz \right. \\ &\quad \left. - \int_0^{c_1} K_h(x - F^{-1}(z))\ dz - \int_{c_m}^1 K_h(x - F^{-1}(z)) dz \right| \end{aligned} \tag{4.4}$$

$$\leq \left| \sum_{j=1}^{m-1} \int_{c_j}^{c_{j+1}} K_h(x - F^{-1}(j/m)) - K_h(x - F^{-1}(z))dz \right| + \left| \int_0^{c_1} K_h(x - F^{-1}(z))dz + \int_{c_m}^1 K_h(x - F^{-1}(z))dz \right|$$

を得る．付録 B で示すように，(4.4)の最後の表現の第 1 項は

$$\left| \sum_{j=1}^{m-1} \int_{c_j}^{c_{j+1}} K_h(x - F^{-1}(j/m)) - K_h(x - F^{-1}(z))dz \right| \leq \frac{1}{mh} \int_{-\infty}^{\infty} |K'(u)|du$$

によって，第 2 項は

$$\left| \int_0^{c_1} K_h(x - F^{-1}(z))dz + \int_{c_m}^1 K_h(x - F^{-1}(z))dz \right| \leq \frac{1}{mh} \sup|K|$$

と押さえられる．従って，カーネル関数 K が絶対連続であり

$$\int |K'(u)|du < \infty, \quad \sup|K| < \infty$$

を満たすならば，x に関わらず一様に $(mh)^{-1}$ のオーダーの量で抑えることができる．言い換えれば，バンド幅 h に比べて，分割数 m が十分大きいならば，パーセント点カーネル推定量と通常のカーネル推定量の差は小さくなる．

よく知られているように，2 次カーネル推定量のバイアスは $\mathrm{E}[\widehat{f}(x|h)] - f(h) = O(h^2)$ である(例えば，Scott, 1992)から，$mh^3 \to \infty$ であれば

$$\mathrm{E}[\widetilde{f}_{\%}(x|h,m)] - f(x) = \mathrm{E}[\widetilde{f}_{\%}(x|h,m)] - \mathrm{E}[\widehat{f}(x|h)] + \mathrm{E}[\widehat{f}(x|h)] - f(h) = \mathrm{E}[\widehat{f}(x|h)] - f(h) + o(h^2)$$

となり，パーセント点カーネル推定量のバイアスは通常のカーネル推定量のバイアスと漸近的に一致する．

4.3 シミュレーションによる例示

4.2 節の考察は，母集団パーセント点が既知であるという非現実的な仮定のもとでなされた．この点を補うために，2.3 節で言及したシミュレーションと同一の状況の下で，パーセント点カーネル推定量の MISE を，通常のヒストグラム及び通常のビン化カーネル推定量の MISE と比較して計算を行った．

4.3.1 通常のヒストグラムに対する効率性

図 5 と図 6 は，パーセント点カーネル推定量(点線)と通常のヒストグラム(連続線)の MISE を比較している．パーセント点カーネル推定量のカーネル関数として，正規カーネル $K(u) = (1/\sqrt{2\pi})e^{-u^2/2}$ を採用し，バンド幅は，スコットのルール(Scott, 1992, (6.17))により，$h = 1.06sn^{-1/5}$ と決めた．ここで，s は標本標準偏差である．

両方のグラフから，m が小さいときには，大きな期待は改善できないが，m が大きくなるにつれ，相当な改善が期待できる．特に，$n = 100$ のときは，$m \geq 7$ の範囲でパーセント点カーネル推定量の方が通常のヒストグラムより効率性が高い．図 6 からは，n が大きくなると，m が小さくない限り，両者の効率性はほぼ等しくなっていくことが示唆される．

4.3.2 通常のビン化カーネル推定量に対する効率性

図 7 と図 8 は，パーセント点カーネル推定量(点線)と通常のビン化カーネル推定量(連続線)を比較している．2.3 節で見たヒストグラム法の場合と比較すると，集計の相違は MISE の大きさに与える影響は少ないように思われる．

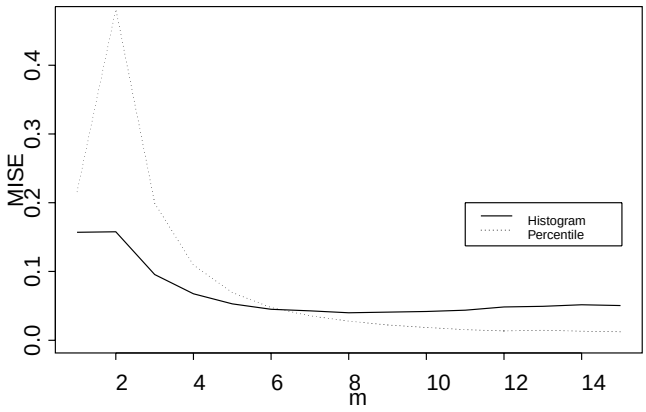

図 5. $n = 100$ **に対する** MISE **の比較：パーセント点カーネル推定量（点線）**VS. **通常のヒストグラム（連続線）.**

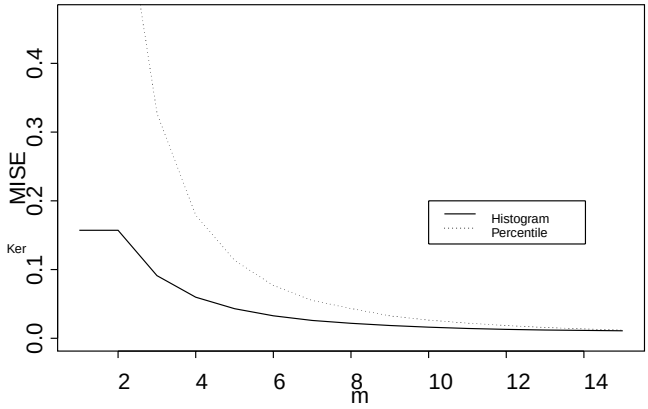

図 6. $n = 1000$ **に対する** MISE **の比較：パーセント点カーネル推定量（点線）**VS. **通常のヒストグラム（連続線）.**

5. 結語

本稿では，先行研究によって示唆されていたパーセント点ヒストグラムにおける理論的問題点—バイアス・パズル—が，実務でしばしば利用される等度数分割による度数表示に対して深刻な影響を与えることを指摘し，バイアス・パズルの考察を行った．カーネル法による改良案を提案し，その有効性を漸近理論によって確かめるとともに，シミュレーションによって例示した.

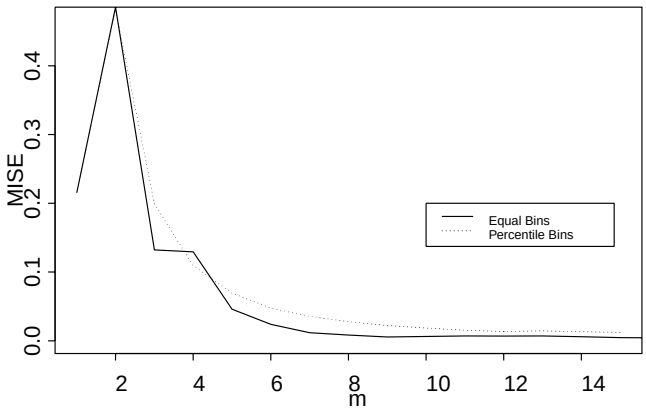

図 7. $n = 100$ **に対する** MISE **の比較：パーセント点カーネル推定量（点線）**VS. **ビン化カーネル推定量.**

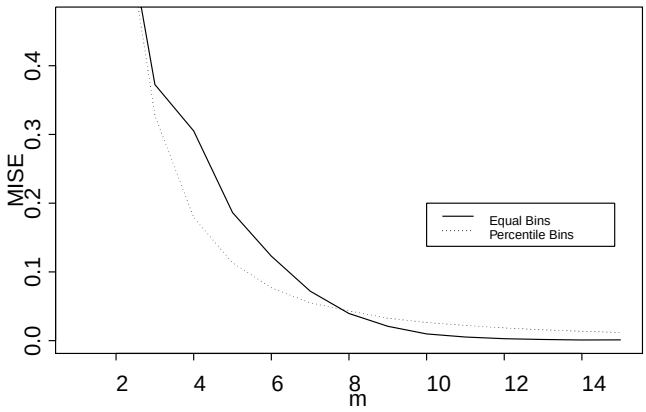

図 8. $n = 1000$ **に対する** MISE **の比較：パーセント点カーネル推定量（点線）**VS. **ビン化カーネル推定量.**

本稿では，既にデータがグループ化されていると言う立場から密度関数推定問題を考察した．これに対して，たとえ原データが利用可能であっても，計算効率性の立場からグルーピングを行うアプローチも利用されている．代表的な手法はヒストグラムである．よく知られているように，ヒストグラム法はカーネル法に比べて理論的な収束の特性が劣る．しかし，近年の研究動向からグループ化データに基づいた推定でも，カーネル法に相当する理論的効率性が実現できることが明らかになってきた．詳細については，Minnotte **(**1996, 1998**)**，Sagae and Scott **(**1997**)**, Sagae and Kogure **(**2000**)****などを見られたい.**

付録 A

$$\begin{aligned}\mathrm{ISB} &= \int_0^1 \left(\mathrm{E}[H_n(x)] - f(x)\right)^2 dx \\ &= \int_0^1 f(x)^2 dx - \frac{1}{m^2}\sum_{j=1}^{m} \frac{1}{\sqrt{j/m} - \sqrt{(j-1)/m}} \\ &= \frac{4}{3} - m^{-3/2}\sum_{j=1}^{m}(\sqrt{j} + \sqrt{j-1}).\end{aligned}$$

Euler-Maclaurin の公式(Lange, 1999, の Proposition 16.2.1 を参照されたい)より

$$\begin{aligned}\sum_{j=1}^{m}\sqrt{j} &= \int_1^m \sqrt{x}dx + \frac{1}{2}\left[m^{1/2}+1\right] + \frac{1}{24}\left[m^{-1/2}-1\right] \\ &\quad -\frac{1}{1920}\left[m^{-5/2}-1\right] + \frac{5}{128}\int_1^m b_4(x)x^{-7/2}dx \\ &= \frac{2}{3}m^{3/2} + \frac{1}{2}m^{1/2} + \frac{\gamma}{2} + \frac{1}{24}m^{-1/2} - \frac{1}{1920}m^{-5/2} - \frac{5}{128}\int_m^\infty b_4(x)x^{-7/2}dx \\ &= \frac{2}{3}m^{3/2} + \frac{1}{2}m^{1/2} + \frac{\gamma}{2} + O(m^{-1/2}).\end{aligned}$$

従って,

$$m^{-3/2}\sum_{j=1}^{m}(\sqrt{j} + \sqrt{j-1}) = \frac{4}{3} - m^{-3/2}\gamma + O(m^2).$$

付録 B

Freedman and Diaconis (1981)の Lemma 2.22 を適用すると，各 j に対して

$$\int_{c_j}^{c_{j+1}} K_h(x - F^{-1}(j/m)) - K_h(x - F^{-1}(z))dz \le \frac{1}{m}\int_{c_{j-1}}^{c_j}\left|K_h'(x - F^{-1}(z))\frac{1}{f(F^{-1}(z))}\right|dz$$

となるから, (4.4)式の最後の表現の第 1 項は

$$\begin{aligned}&\left|\sum_{j=1}^{m-1}\int_{c_j}^{c_{j+1}} K_h(x - F^{-1}(j/m)) - K_h(x - F^{-1}(z))dz\right| \\ &\le \frac{1}{m}\sum_{j=1}^{m-1}\int_{c_{j-1}}^{c_j}\left|K_h'(x - F^{-1}(z))\frac{1}{f(F^{-1}(z))}\right|dz \\ &= \frac{1}{mh}\int_0^1 \frac{1}{h}\left|K'\left(\frac{x - F^{-1}(z)}{h}\right)\frac{1}{f(F^{-1}(z))}\right|dz = \frac{1}{mh}\int_{-\infty}^{\infty}|K'(u)|du\end{aligned}$$

となる．また，右辺の第 2 項は

$$\begin{aligned}&\left|\int_0^{c_1} K_h(x - F^{-1}(z))dz + \int_{c_m}^1 K_h(x - F^{-1}(z))dz\right| \\ &= \left|\int_{-\infty}^{F^{-1}(1/2m)} K_h(x-y)f(y)dy + \int_{F^{-1}(1-1/2m)}^{\infty} K_h(x-y)f(y)dy\right| \\ &\le \frac{\sup|K|}{h}\left|\int_{-\infty}^{F^{-1}(1/2m)} f(y)dy + \int_{F^{-1}(1-1/2m)}^{\infty} f(y)dy\right| \le \frac{1}{mh}\sup|K|\end{aligned}$$

となる.

参 考 文 献

Fan, J. and Marron, J. S. (1994). Fast implementation of nonparametric curve estimators, *Journal of Computational and Graphical Statistics*, **3,** 35–56.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory, *Zeitschrift für Wahrscheinlichkeitstheorie und Vervandte Gebiete*, **57**, 453–476.

Hall, P. and Wand, M. P. (1996). On the accuracy of binned kernel density estimators, *Journal of Multivariate Analysis*, **56**, 165–184.

Jones, M. C. (1991). The roles of ISE and MISE in density estimation, *Statistics & Probability Letters*, **12**, 51–56.

Kogure, A. (1987). Asymptotically optimal cells for a histogram, *Annals of Statistics*, **15**, 1023–1030.

Lange, K. (1999). *Numerical Analysis for Statisticians*, Springer, New York.

Lecoutre, J. P. (1987). The histogram with random partition, *New Perspectives in Theoretical and Applied Statistics*, 265–276, John Wiley and Sons, New York.

Minnotte, M. C. (1996). The bias-optimized frequency polygon, *Computational Statistics*, **11**, 35–48.

Minnotte, M. C. (1998). Achieving higher-order convergence rates for density estimation with binned data, *Journal of the American Statistical Association*, **93**, 663-672.

Sagae, M. and Kogure, A. (2002). Maximum entropy density estimator, *Proceedings of JSM2002, Joint Statistical Meetings 2002, New York*.

Sagae, M. and Scott, D. W. (1997). Bin interval method of locally nonparametric density estimation, Tech. Report, Department of Statistics, Rice University, Houston, Texas.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York.

Terrell, G. R. and Scott, D. W. (1992). Variable Kernel density estimation, *Annals of Statistics*, **20**, 1236–1265.

Proceedings of the Institute of Statistical Mathematics Vol. 53, No. 2, 375–389 (2005) 389

Estimating Probability Density from Percentiles

Atsuyuki Kogure[1] and Masahiko Sagae[2]

[1]Faculty of Policy Management, Keio University
[2]Faculty of Engineering, Gifu University

We consider the problem of estimating density from percentiles. Prior results such as Lecoutre (1987) and Scott (1992) suggest that a histogram with percentiles suffers large biases. We re-examine the bias problem and show that the kernel method alleviates it by asymptotic arguments. To confirm the theoretical results we give simple simulation studies.

Key words: Histogram, percentiles, bin frequency, kernel estimator, bias.