Kanazawa University,
Faculty of Economics and Management
Discussion Paper Series
No. 070

有限区間におけるHistogramのビン幅補正について

齊藤実祥
寒河江雅彦

saito_misaki@stu.kanazawa-u.ac.jp
sagae.masahiko@gmail.com

2 October 2022

金沢大学経済学経営学系
〒920-1192 金沢市角間町
Faculty of Economics and Management,
Kanazawa University
Kakumamachi, Kanazawa-shi, Ishikawa, 920-1192, Japan
https://keikei.w3.kanazawa-u.ac.jp/research_dp.html

有限区間における Histogram のビン幅補正について

齊藤実祥 (金沢大学 人間社会環境研究科)
寒河江雅彦 (金沢大学 人間社会研究域)

Bin-width Correction of Histogram in a Finite Interval

Misaki Saito (Kanazawa University)
Masahiko Sagae (Kanazawa University)

定義域が与えられている Histogram の始点を定義域の最小値にし、ビン幅を推定して構築した Histogram の推定区間は一般に定義域と一致しない。この推定した区間と定義域のずれ (「ビン残差」と呼ぶ) を解消する単純な方法は、ビン残差を各ビンに等分配してビン幅を補正することである。本稿では、この補正法の漸近的性質と有限標本における特性を示し、補正の有用性について理論面から明らかにする。また、数値実験によるその有効性を確かめる。単純な補正法であるため、様々なビン幅推定法でも適用可能な手法と言える。

Given an interval, a range of Histogram constructed by starting the Histogram at the minimum value of the interval and estimating a bin width generally does not match the interval. A simple way to eliminate the gap between the estimate and the interval (called the "bin residuals") is to divide the bin residual evenly into each bins and correct the bin width. In this paper, we show the asymptotic properties of this correction and its characteristics in finite samples, and clarify the usefulness of the correction from a theoretical aspect. We also examine its effectiveness by numerical experiments. Since it is a simple correction method, it can be applied various bin-width estimation methods.

1 はじめに

本稿では、閉区間の定義域が与えられている場合を考える。Histogram の始点を定義域の最小値に定め、ビン幅を推定後に構築した Histogram の推定区間は一般に定義域と一致しない点に着目した。この推定区間と定義域とのずれについて、以降では「ビン残差」と呼ぶ。ビン残差に関わる問題は、一般的によく用いられる計算・統計処理ソフトウェアでも見られる。例えば、R の hist 関数におけるデフォルトの設定では、スタージェスのルールで決定されるビン数を用いて Histogram が構築される。スタージェスのルールとは、Sturges(1926) の提案した非常に単純なビン数の決定法のため、ビン残差の問題は生じない。しかしながら、Scott(1979) はスタージェスのルールが理論的根拠に乏しい決定手法であることを指摘し、平均積分二乗誤差 (MISE) 基準に基づいたビン幅を推定する方法を提案しており、スコットのルールとして広く用いられている。他の例として、Excel における Histogram のデフォルトの設定では、Histogram の始点をデータの最小値とし、スコットのルールでビン幅が推定されるが、ビン残差が生じて定義域と異なる推定区間での Histogram が構築される。

ビン残差の問題を解消するためには、推定区間と定義域が一致するようにビン幅を補正する必要がある。ビン幅の補正法で最も単純なものは、ビン残差をビン数で割って各ビンに等分配することである。換言すると、定義域 $[a,\ b]$、何らかの方法で決められたビン幅 h^* とすると、補正後のビン幅 $\tilde{h}$ とは $\tilde{h} = (b-a)/\left[\frac{b-a}{h^*} + \frac{1}{2}\right]$ と書ける。ここで [ ] はガウス記号である。この補正法が Histogram 推定に対してどのような影響を及ぼすかについて、その理論面は未整備である。したがって、本稿では、既知の定義域上のビン幅補正後の Histogram 推定量について、MISE 基準に基づいた漸近的性質を導出し、数値実験からこの補正法の有用性について考察する。

2 Histogram の定義及び漸近的性質

この章では、ビン幅の補正に際して必要となる Histogram の定義及びその MISE 基準に基づいた漸近的性質について説明する。

2.1 Histogram の定義

確率密度関数 $f(x)$ からの n 個のデータを $\{X_1, X_2, \ldots, X_n\}$、$k$ 番目のビンを B_k、ビン B_k の区間 $[t_{k-1},\ t_k)$、ビン幅 h、ビン B_k に入る度数を ν_k とする。このとき、Histogram は次式で与えられる;

$$\hat{f}(x;h) = \frac{\nu_k}{nh} = \frac{1}{nh}\sum_{i=1}^{n} I_{[t_{k-1},\ t_k)}(X_i)\ ,\ x \in B_k, \tag{1}$$

ただし、$I_A(x)$ は定義関数で以下の通りに定義される;

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

2.2 Histogram の漸近的性質

推定モデルの良さを測る評価基準として、推定量 $\hat{f}$ と真の密度関数 f との MISE を用いる。MISE の定義は次の通りである;

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

ここで、IV は積分分散、ISB は積分二乗バイアスに対応する。

(1) で定義された Histogram の漸近的な MISE(AMISE) は、

$$\begin{aligned} \mathrm{AMISE}_{HIST}[\hat{f}(x;h)] &= \mathrm{AIV}_{HIST} + \mathrm{AISB}_{HIST} \\ &= \frac{1}{nh} + \frac{R(f')}{12}h^2, \end{aligned} \tag{3}$$

ただし、$R(f') = \int f'(x)^2 dx$ である。このとき、最適ビン幅 h^* と最小 AMISE$[\hat{f}(x;h^*)]$ は、次式で与えられる;

$$h^* = \left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}, \tag{4}$$

$$\text{最小 AMISE}[\hat{f}(x;h^*)] = \frac{3}{2}\left(\frac{R(f')}{6}\right)^{\frac{1}{3}} n^{-\frac{2}{3}}. \tag{5}$$

$R(f')$ は真の分布 $f(x)$ に依存する未知の定数とみなすことができるため、Histogram の精度はデータ数 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。

3　ビン幅補正法

3.1　ビン幅の補正法について

本稿では、Histogram の定義域は既知の区間とする。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。この条件の下で、n 個のデータが与えられたとき、(4) より AMISE$[\hat{f}(x;h)]$ 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^* はデータ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、

$$m = \frac{b-a}{h^*} + \frac{1}{2}, \tag{6}$$

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。

$$m^* = \left[\frac{b-a}{h^*} + \frac{1}{2}\right], \tag{7}$$

$$\eta := [0,\ 1), \tag{8}$$

ただし、m^* はガウス記号を用いて表現している。

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、確率密度関数 $g(\delta)$ に従うとものする。こ

のとき、δ は以下のように表される;

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

$$-\frac{h^*}{2} \leq E[\delta] \leq \frac{h^*}{2}, \tag{9}$$

$$0 \leq E[\delta^2] \leq \frac{h^{*2}}{4}, \tag{10}$$

$$0 \leq \mathrm{Var}[\delta] \leq \frac{h^{*2}}{4}, \tag{11}$$

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、ビン残差 δ をビン数 m^* で等分割し、ビン幅 h^* に加える。この補正後のビン幅を $\tilde{h}$ とすると、

$$\tilde{h} = h^* - \frac{\delta}{m^*}, \tag{12}$$

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

$$\hat{f}(x; \tilde{h}) = \frac{\tilde{\nu}_k}{n\tilde{h}} = \frac{\tilde{\nu}_k}{n} \frac{m^*}{m^* h^* - \delta}, \tag{13}$$

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

3.2 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。

この条件のもとで、以下の定理が成り立つ。ここで、h^* は AMISE$[\hat{f}(x; h)]$ 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

$h \propto O(n^{-\alpha}), x \in B_k$ に対して、
$\alpha = \frac{1}{3}$ のとき、

$$\sqrt{nh^*}\left\{\hat{f}(x;\tilde{h}) - f(x)\right\} \xrightarrow{d} N\left(\mathrm{Bias}[\hat{f}(x;\tilde{h})], f(\xi_k)\left(1 + \frac{\delta}{m^*h^*}\right)\right), \tag{14}$$

$\alpha > \frac{1}{3}$ のとき、

$$\sqrt{nh^*}\left\{\hat{f}(x;\tilde{h}) - f(x)\right\} \xrightarrow{d} N\left(o(1), f(\xi_k)\left(1 + \frac{\delta}{m^*h^*}\right)\right), \tag{15}$$

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

ビン幅補正後 Histogram の漸近一致性の証明は Appendix 1、漸近正規性の証明は Appendix 2 で詳述する。

3.3 MISE の上限と下限

この節では、補正後ビン幅を用いた Histogram の $E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right]$ の上限と下限をビン残差 δ について場合分けして示す。

3.3.1 $E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right]$ 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、

$$\begin{aligned} E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] = E_\delta\left[n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] + E_\delta\left[\frac{R(f')}{12}h^{*2}\middle|h^*\right] \\ - E_\delta\left[\frac{R(f')}{6}\frac{\delta}{m^*}h^*\middle|h^*\right] + E_\delta\left[\frac{R(f')}{12}\frac{\delta^2}{m^{*2}}\middle|h^*\right]. \end{aligned} \tag{16}$$

(16) 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、(16) 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 (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 $\delta = \pm\frac{h^*}{2}$ での左側一点分布、右側一点分布または両端二点対称分布の時に上限値を取る。したがって、$-\frac{h^*}{2} \le E[\delta] \le \frac{h^*}{2},\ 0 \le E[\delta^2] \le \frac{h^{*2}}{4}$ を (16) に代入して、

$$\frac{1}{nh^*} + \frac{R(f')}{12}h^{*2} \le E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \le \frac{1}{nh^*} + \frac{R(f')}{12}h^{*2} + \frac{R(f')}{48m^{*2}}h^{*2}. \tag{17}$$

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

$$\mathrm{AMISE}[\hat{f}(x;h^*)] \le E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \le \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{R(f')}{48m^{*2}}h^{*2}, \tag{18}$$

となる。

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

$$\mathrm{AMISE}[\hat{f}(x;h^*)] \leq E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \leq \mathrm{AMISE}[\hat{f}(x;h^*)] + O(n^{-\frac{4}{3}}). \tag{19}$$

3.3.2 ビン残差 δ の分散が最大になる場合

ビン残差 δ の分散が最大となるのは、δ が $-\frac{h^*}{2}$ と $\frac{h^*}{2}$ を取る両側二点対称分布の場合である。したがって、$-\frac{h^*}{2} \leq E[\delta] \leq \frac{h^*}{2},\ 0 \leq E[\delta^2] \leq \frac{h^{*2}}{4}$ を (16) に代入して、

$$\mathrm{AMISE}[\hat{f}(x;h^*)] \leq E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \leq \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{R(f')}{48m^{*2}}h^{*2}. \tag{20}$$

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

$$\mathrm{AMISE}[\hat{f}(x;h^*)] \leq E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \leq \mathrm{AMISE}[\hat{f}(x;h^*)] + O(n^{-\frac{4}{3}}). \tag{21}$$

3.3.3 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、$-\frac{h^*}{2} \leq E[\delta] \leq \frac{h^*}{2},\ 0 \leq E[\delta^2] \leq \frac{h^{*2}}{12}$ であるため (16) に代入して、

$$\frac{1}{nh^*} + \frac{R(f')}{12}h^{*2} \leq E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \leq \frac{1}{nh^*} + \frac{R(f')}{12}h^{*2} + \frac{R(f')}{144m^{*2}}h^{*2}. \tag{22}$$

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

$$\mathrm{AMISE}[\hat{f}(x;h^*)] \leq E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \leq \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{R(f')}{144m^{*2}}h^{*2}. \tag{23}$$

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

$$\mathrm{AMISE}[\hat{f}(x;h^*)] \leq E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] \leq \mathrm{AMISE}[\hat{f}(x;h^*)] + O(n^{-\frac{4}{3}}). \tag{24}$$

3.3.1〜3 項の議論から、ビン残差 δ の分布によらず、漸近的には AMISE に影響を与えないことが分かる。

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、まずビン残差 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。

4.1　定義域 $[-3,\ 0], [0,\ 3]$ での数値実験結果

表 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 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。表は小さな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 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 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	<u>0.031301</u>	<u>0.013852</u>	<u>0.008022</u>	<u>0.005208</u>	<u>0.001870</u>

表 3 は定義域 $[-3,\ 0]$ の ISE 標準偏差、表 4 は定義域 $[0,\ 3]$ の ISE 標準偏差の数値実験結果を示している。表中で小さな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 標準偏差の値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 標準偏差の値の方が小さく、分散の安定化に有効であることが分かる。

表 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	<u>0.021346</u>	<u>0.007496</u>	<u>0.003598</u>	<u>0.002119</u>	<u>0.000587</u>

表 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

4.2　定義域 $[-3,\ 3]$ での数値実験結果

表 5 は補正なし最適ビン幅と補正後ビン幅の MISE、表 6 は補正なし最適ビン幅と補正後ビン幅の ISE 標準偏差の数値実験結果を示している。表は小さな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 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 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。他方、ビン幅補正は分散安定化に有効である。

表 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 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

4.3　定義域 $[-1,\ 1]$ での数値実験結果

表 7 は補正なし最適ビン幅と補正後ビン幅の MISE、表 8 は補正なし最適ビン幅と補正後ビン幅の ISE 標準偏差の数値実験結果を示している。表は小さな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 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 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 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 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　結論と考察

閉区間の定義域が与えられた場合に、ビン幅を推定して Histogram 推定を行うとビン残差が生じ、定義域と推定区間のずれが生じる問題に着目した。定義域を満たすための簡単な方法として、ビン残差を各ビンに等分配してビン幅を補正することが考えられる。理論面に関し未整備であったため、ビン幅補正後の Histogram について漸近的性質を導出し、数値実験から有限標本特性を調べることで、ビン幅補正の有用性について確かめた。

ビン幅補正の漸近的性質について、$E[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]]$ は $\mathrm{AMISE}[\hat{f}(x;h^*)]$ に漸近的に一致することを示し、$\hat{f}(x;\tilde{h})-f(x)$ は漸近正規性が成り立つことを示した。また、$E[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]]$ の上限と下限を導出した。

有限標本でのビン幅補正による特性を確かめるため MISE の数値実験を行った。その結果、tail 部分の確率が大きい場合、すなわち、ビン残差部分にデータが多い場合には MISE への影響が大きいことが明らかになった。また、tail 部分の確率が大きい場合には、データ数が多い場合であっても補正後ビン幅の MISE の方が小さく、分散も安定化することから、補正が有効に働くと考えられる。

ビン幅の推定法には、スコットのルールをはじめ様々な方法があるが、本稿で議論したビン幅補正法は、様々なビン幅選択法であっても適用可能である。

Appendix 1

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。$\mathrm{AMISE}_{HIST}[\hat{f}(x;h)]=\frac{1}{nh}+\frac{R(f')}{12}h^2$ に

$\hat{f}(x;\tilde{h})$ を代入して、

$$\begin{aligned}\mathrm{AMISE}[\hat{f}(x;\tilde{h})] &= \frac{1}{n\tilde{h}} + \frac{1}{12}\tilde{h}^2 R(f') \\ &= \frac{1}{n\left(h^* - \frac{\delta}{m}\right)} + \frac{1}{12}\left(h^* - \frac{\delta}{m}\right)^2 R(f') \\ &= n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1} + \frac{R(f')}{12}h^* - \frac{R(f')}{6}\frac{\delta}{m^*}h^* + \frac{R(f')}{12}\frac{\delta^2}{m^{*2}}. \end{aligned} \tag{25}$$

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、

$$\begin{aligned}E_\delta\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\middle|h^*\right] = E_\delta\left[n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] + E_\delta\left[\frac{R(f')}{12}h^{*2}\middle|h^*\right] \\ - E_\delta\left[\frac{R(f')}{6}\frac{\delta}{m^*}h^*\middle|h^*\right] + E_\delta\left[\frac{R(f')}{12}\frac{\delta^2}{m^{*2}}\middle|h^*\right]. \end{aligned} \tag{26}$$

(26) の第 1 項について、

$$\begin{aligned}E_\delta\left[n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] &= n^{-1}E_\delta\left[\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] \\ &= \frac{m^*}{n}E_\delta\left[\frac{1}{m^*h^* - \delta}\middle|h^*\right] \\ &= \frac{m^*}{n}\int_{-\frac{h^*}{2}}^{\frac{h^*}{2}}\frac{1}{m^*h^* - \delta}\frac{1}{h^*}d\delta \\ &= \frac{m^*}{n}\frac{1}{h^*}\left[\log|m^*h^* - \delta|\right]_{-\frac{h^*}{2}}^{\frac{h^*}{2}} \\ &= \frac{m^*}{n}\frac{1}{h^*}\log\left(\frac{m^*h^* + \frac{h^*}{2}}{m^*h^* - \frac{h^*}{2}}\right) \\ &= \frac{m^*}{n}\frac{1}{h^*}\log\left(\frac{1 + \frac{1}{2m^*}}{1 - \frac{1}{2m^*}}\right). \end{aligned} \tag{27}$$

(27) 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、

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

と近似される。

したがって、(26) の第 1 項は以下の式で近似される;

$$E_\delta\left[n^{-1}\left(h^* - \frac{\delta}{m^*}\right)^{-1}\middle|h^*\right] \approx \frac{1}{nh^*}\left(1 + \frac{1}{3}\left(\frac{1}{2m^*}\right)^2\right). \tag{28}$$

(26) の第 2 項は、δ を含まない項のため定数となり、

$$E_\delta\left[\frac{R(f')}{12}h^{*2}\middle|h^*\right] = \frac{R(f')}{12}h^{*2}. \tag{29}$$

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

$$E_\delta\left[\frac{R(f')}{6}\frac{\delta}{m^*}h^*\middle|h^*\right] = 0. \tag{30}$$

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

$$E_\delta\left[\frac{R(f')}{12}\frac{\delta^2}{m^{*2}}\middle|h^*\right] = \frac{R(f')}{12m^{*2}}E_\delta[\delta^2] = \left(\frac{h^*}{12m^*}\right)^2 R(f'). \tag{31}$$

以上 (28)〜(31) 式より、$E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right]$ は、以下の通りである;

$$\begin{aligned} E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right] &\approx \frac{1}{nh^*} + \frac{1}{nh^*}\left(\frac{1}{12m^{*2}}\right) + \frac{R(f')}{12}h^{*2} + \frac{1}{12m^{*2}}\frac{R(f')}{12}h^{*2} \\ &= \left(1 + \frac{1}{12m^{*2}}\right)\mathrm{AMISE}[\hat{f}(x;h^*)]. \end{aligned} \tag{32}$$

(32) 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、

$$\begin{aligned} \frac{1}{12m^{*2}} &= \frac{1}{12}\left\{\frac{b-a}{h^*} - \left(\eta - \frac{1}{2}\right)\right\}^{-2} \\ &= \frac{h^{*2}}{12\left\{(b-a) - \left(\eta - \frac{1}{2}\right)h^*\right\}^2} \\ &= \frac{\left(\frac{6}{R(f')}\right)^{\frac{2}{3}} n^{-\frac{2}{3}}}{12\left\{(b-a) - \left(\eta - \frac{1}{2}\right)\left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}\right\}^2}, \end{aligned} \tag{33}$$

となる。

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

$$\begin{aligned} E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})]\right] &= \left(1 + \frac{1}{12m^{*2}}\right)\mathrm{AMISE}[\hat{f}(x;h^*)] \\ &= \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{\left(\frac{6}{R(f')}\right)^{\frac{2}{3}} n^{-\frac{2}{3}}}{12\left\{(b-a) - \left(\eta - \frac{1}{2}\right)\left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}\right\}^2}\mathrm{AMISE}[\hat{f}(x;h^*)]. \end{aligned} \tag{34}$$

(34) 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、

$$E\left[\mathrm{AMISE}[\hat{f}(x;\tilde{h})\right]] = \mathrm{AMISE}[\hat{f}(x;h^*)] + \frac{\left(\frac{6}{R(f')}\right)^{\frac{1}{3}}}{8\left\{(b-a) - \left(\eta - \frac{1}{2}\right)\left(\frac{6}{R(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}\right\}^2} n^{-\frac{4}{3}}. \tag{35}$$

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 (35) の第 2 項を加えたものとなる。また、(35) 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、(35) で $h^* \propto n^{-\frac{1}{3}}$, AMISE$[\hat{f}(x;h^*)] = O\left(n^{-\frac{2}{3}}\right)$ より、

$$\begin{aligned} E\left[\text{AMISE}[\hat{f}(x;\tilde{h})]\right] &= \text{AMISE}[\hat{f}(x;h^*)] + O\left(n^{-\frac{4}{3}}\right) \\ &= \text{AMISE}[\hat{f}(x;h^*)] + o\left(n^{-\frac{2}{3}}\right), \end{aligned} \tag{36}$$

であることから、$n \to \infty$ のとき、

$$E\left[\text{AMISE}[\hat{f}(x;\tilde{h})]\right] \xrightarrow{d} \text{AMISE}[\hat{f}(x;h^*)], \tag{37}$$

となり、

$$\int \left(\hat{f}(x;\tilde{h}) - f(x)\right)^2 dx \xrightarrow{d} 0, \tag{38}$$

であることから、

$$\hat{f}(x;\tilde{h}) \xrightarrow{d} f(x), \tag{39}$$

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。

Appendix 2

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_k における $\hat{f}(x;\tilde{h})$ は以下の通りである。

$$\hat{f}(x;\tilde{h}) = \frac{\tilde{\nu}_k}{n\tilde{h}} = \frac{\tilde{\nu}_k}{n}\frac{m^*}{m^*h^* - \delta}. \tag{40}$$

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

$$E\left[\hat{f}(x;\tilde{h})\right] = \frac{m^*}{n\left(m^*h^* - \delta\right)} E\left[\tilde{\nu}_k\right] = \frac{m^*}{m^*h^* - \delta}\tilde{p}_k, \tag{41}$$

$$\begin{aligned} \text{Var}\left[\hat{f}(x;\tilde{h})\right] &= \frac{1}{\left(n\tilde{h}\right)^2}\text{Var}\left[\tilde{\nu}_k\right] \\ &= \frac{\tilde{p}_k\left(1-\tilde{p}_k\right)}{n\left(h^* - \frac{\delta}{m^*}\right)^2}. \end{aligned} \tag{42}$$

$\tilde{p}_k$ について積分の平均値の定理より、

$$\tilde{p}_k = \int_{B_k} f(t)dt = \tilde{h}f(\xi_k) = \left(h^* - \frac{\delta}{m^*}\right) f(\xi_k), \tag{43}$$

ただし、ξ_k は $\xi_k \in B_k$ を満たす B_j 内のある点とする。

(42) に (43) を代入すると、

$$\begin{aligned}
\mathrm{Var}\left[\hat{f}(x;\tilde{h})\right] &= \frac{\left(h^* - \frac{\delta}{m^*}\right) f(\xi_k) \left\{1 - \left(h^* - \frac{\delta}{m^*}\right) f(\xi_k)\right\}}{n\left(h^* - \frac{\delta}{m^*}\right)^2} \\
&= \frac{f(\xi_k)}{n\left(h^* - \frac{\delta}{m^*}\right)} + O\left(n^{-1}\right) \\
&\approx \frac{m^* f(\xi_k)}{n} \frac{1}{m^* h^* - \delta} \\
&\approx \frac{f(\xi_k)}{nh^*}\left(1 + \frac{\delta}{m^* h^*}\right). \qquad (44)
\end{aligned}$$

したがって、二項分布の中心極限定理と (44) を用いることで、次の通りの表現を得る。

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

$\alpha = \frac{1}{3}$ のとき、

$$\sqrt{nh^*}\left\{\hat{f}(x;\tilde{h}) - f(x)\right\} \xrightarrow{d} N\left(\mathrm{Bias}[\hat{f}(x;\tilde{h})],\ f(\xi_k)\left(1 + \frac{\delta}{m^* h^*}\right)\right), \qquad (45)$$

$\alpha > \frac{1}{3}$ のとき、

$$\sqrt{nh^*}\left\{\hat{f}(x;\tilde{h}) - f(x)\right\} \xrightarrow{d} N\left(o(1),\ f(\xi_k)\left(1 + \frac{\delta}{m^* h^*}\right)\right), \qquad (46)$$

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。

参考文献

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

[2] 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-pp.476.

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

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

[5] 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.

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

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

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