人間社会環境研究　第42号　2021. 9

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

人間社会環境研究科　人間社会環境学専攻
齊　藤　実　祥
人間社会研究域　経済学経営学系
寒河江　雅　彦

要旨

ヒストグラムはデータの取りうる範囲を重複しない区間（ビン）に分け，その各区間内に入るデータ数によって決まる分布の推定法である。ビンは分割点を決めることで定まり，平均積分二乗誤差基準をもとにビン幅推定の理論が構成される。等間隔ビンの推定法はScott（1979）で提案され，不等間隔のビンに関してはKogure（1987）以降，多くの提案がなされてきた。その中で，Lecoutre（1987）は，分割点をデータから等パーセンタイルで選択する手法を提案し，その漸近的性質を導出している。ヒストグラムの改良について，Scott（1985）は等間隔のビンをシフトさせながらヒストグラムを推定するAveraged Shifted Histogram（ASH）を提案した。我々は，ヒストグラムの分割点を一様乱数により決定して不等間隔のヒストグラムを推定し，これを繰り返して得られたヒストグラムの平均を推定量とするRandom Partitioned Histogram（以降，RPH）を提案する。本稿では，ヒストグラムとの比較を通じてRPHの有限サンプルにおける性質と有効性をシミュレーションにより明らかにする。数値実験では，データ数，分割数，繰り返し回数を変化させ，様々なパターンで実験を行い，推定精度にどう影響するかを調べる。

シミュレーションの結果，RPHは分布の形状とサンプル数に関わらず，ビン数と繰り返し回数の適切な選択をする限りヒストグラムと比較してISE値が減少し，分散も安定化して推定精度は改良されることが明らかになった。RPHのビン数については，ヒストグラムの最適ビン数を超えて設定した場合，より推定精度が優れていることが分かった。また，サンプル数が小さい場合には，RPH推定時の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，サンプル数が大きい場合には，繰り返し回数が少ない場合でも十分に推定精度は高い。ヒストグラムでは最適でないビン幅を選択すると平滑化不足または平滑化過多により推定精度は悪くなるが，RPHでは平滑化不足または平滑化過多であっても，多くの場合ヒストグラムよりも推定精度が改良されることが分かった。したがって，RPHはデータに依存するが，分布の形状に関わらず，適用範囲が広く有効な推定手法であることが示された。

キーワード

ヒストグラム，Random Partitioned Histogram，グループ数

Simulation Study for Random Partitioned Histogram

Division of Human and Socio-Environmental Studies
Graduate School of Human and Socio-Environmental Studies
SAITO Misaki
Faculty of Economics and Management Institute of Human and Social Sciences
SAGAE Masahiko

Abstract

Researchers have discussed methods for selecting equal or unequal partitions of histograms. Lecoutre（1987）proposed a histogram with the equipercentile bins at data points and derived its asymptotic properties. In several previous studies, the histogram is estimated at once. Scott（1985）proposed the Averaged Shifted Histogram（ASH）, which repeatedly constructs the histogram while shifting the division points at equal bins and uses their average as the estimator. We refer to studies by Lecoutre and Scott and propose an estimator called the Random Partitioned Histogram（RPH）. The RPH is defined to be the equally weighted average of the histogram, that is estimated by using random partitioned bins. Asymptotic properties in large samples and properties of finite samples of RPH have not been shown. Therefore, we examine the properties of RPH in a finite sample by comparing it to a histogram. Simulations are performed using various cases to investigate how the method of selecting the number of bins and repetitions, which are RPH parameters, affects the estimation accuracy.

As a result, the ISE of RPH decreases regardless of the sample size, the variance stabilizes, and the estimation accuracy improves, if the number of bins and iterations are not extremely selected. In addition, it was found that it is desirable to select a large number of RPH bins. When the number of samples is small, it is better to increase the number of iterations for RPH estimation. Meanwhile, when the number of samples is large, the estimation accuracy is high even when the number of iterations is small. When the number of samples is sufficiently large, the estimation accuracy of RPH improves compared to the histogram, even if the number of bins selected is too large.

Keyword

Histogram, Random Partitioned Histogram, Number of Group

１．研究背景と目的

統計的データの代表的なグラフとして，ヒストグラムが挙げられる。ここでヒストグラムとは密度関数を指し，分割された重複しない各区間（以降，ビン）に入る度数データに比例した面積を持つ柱状グラフで構成された連続分布である。

等間隔ヒストグラムのビン幅推定に関しては，Scott（1979）が二乗誤差基準の一つである平均積分二乗誤差（以降，MISE）を根拠とした推定手法を提案した[1)]。この手法では参照分布として正規分布を仮定し，未知のスケールパラメータである標準偏差の代わりにサンプルの標準偏差を用いてビン幅を推定している。また，不等間隔ヒストグラムについてはKogure（1987）やTerrell and Scott（1992）によって議論された。不等間隔ヒストグラムは，等間隔ヒストグラムよりも推定精度が改良されるものの，その改良の程度は小さいことが示されている。この他，Shibuya and Yamato (1995) が確率分布に従う分割点の決定法

を議論している。また，Lecoutre（1987）は，分割点をデータから等パーセンタイルで選択する手法を提案し，漸近的性質を導出している。これらの先行研究はデータセットに対して一回だけで推定する手法である。一方で，Scott（1985）はヒストグラムを繰り返し推定に利用することを考えた。ヒストグラムを構築した後，同じデータに対して分割点をシフトさせたヒストグラムを再構築し，それらの平均を推定量とするAveraged Shifted Histogram（ASH）を提案し，理論的にヒストグラムよりも推定精度が改良されることを示した。

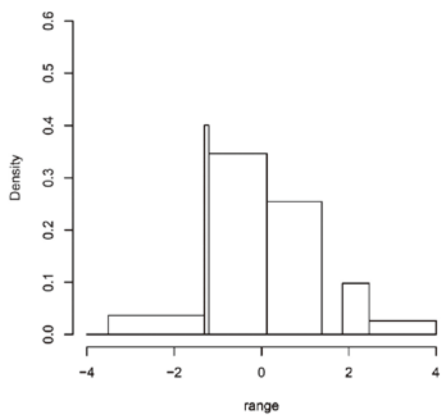
我々はLecoutre（1987）による分割点の選択法とScott（1985）のASHを参考に，ヒストグラムの分割点を一様乱数により決定して不等間隔ヒストグラムを推定し，これを繰り返して得られたヒストグラムの平均を推定量とすることを考える。本稿では，この推定量をRandom Partitioned Histogram（以降，RPH）と呼ぶ。詳細については２章で述べるが，RPHは分割点の決定時に特定の確率分布を想定しないため，従来手法より緩い条件での決定法であり，また，分析者のヒストグラム作成時に決定されるBin Originとビン幅あるいはビン数を厳密に決めることなく推定可能である。RPHについては，大標本における漸近的性質及び有限サンプルにおける性質が未解決な問題である。したがって，本稿では，RPHの性質と大標本特性を数値実験により明らかにすることを目的とする。具体的には，シミュレーション実験をデータ数，ビン数，繰り返し回数を変化させて行い，それらの結果からヒストグラムとRPHとの推定精度についてISEの平均と標準偏差により比較し，RPHの推定における精度と安定性（頑健性）及びその有効性を考察する。

２．Random Partitioned Histogram (RPH)

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・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ヒストグラムを構築する。（ⅲ）（ⅱ）の操作を繰り返し，推定したヒストグラムの平均をRPH推定量とする。以下，各手順について具体的に説明する。

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[2)]やスコットのルール等のビン幅推定法を用いてヒストグラムの分割数を決定する。

手順（ⅱ）で，一般にヒストグラムの始点と終点は未知であるが，本稿では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。（ⅰ）で決定したビン数に対応し



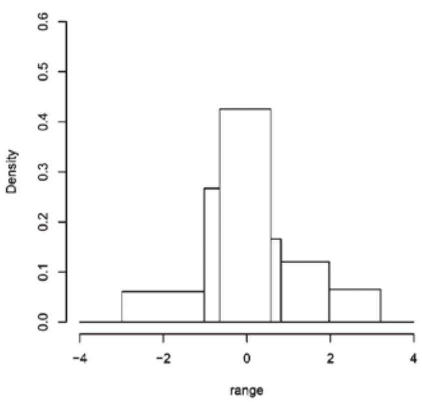


図１　Random Partitionの例（分割数8，繰り返しなし）

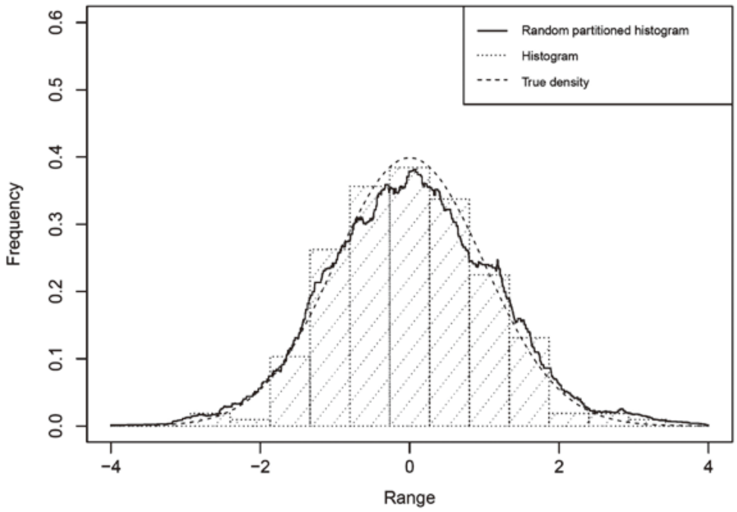


図２　RPH推定量の例

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ヒストグラムの推定を行う。図１は，繰り返しなしのランダムに分割したビンでのヒストグラム推定の例である。定義域［-4, 4］で標準正規分布に従う50個のデータ，ビン数を８として乱数で決定した分割点に基づくヒストグラムの例を示す。両グラフともに同じデータを用いているが，分割点が毎回ランダムに決定されるため，異なるヒストグラムが推定される。また，一般的な不等間隔ヒストグラムと同様に，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。

手順（ⅲ）では，（ⅱ）を繰り返して複数のヒストグラムを構築し，その平均をRPH推定量とする。図２の実線のグラフは，標準正規分布に従う200個のサンプルについて，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ヒストグラムの平均であるため，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，本稿では定義域が与えられたRPHについてシミュレーション計算を行うが，その際には広めの定義域を設定する。

３．シミュレーション設定

ヒストグラムとRPHの密度推定の精度を比較するため，積分二乗誤差（以降，ISEと呼ぶ）について計算シミュレーションを行う。ここでは，MISEの変動を評価するため，ISEの計算シミュレーションを10,000回行い，ISE値とその標準偏差を計算する。以降，シミュレーション10,000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$N(0,1)$に従うサンプルについて，表１で示す通りの設定でシミュレーションを行う。表１でサンプル数をn，ビン数をm，繰り返し回数をrとし，以降も同様とする。

全てのシミュレーションにおいて，ヒストグラム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，各シミュレーションの詳細を説明する。

ケース①：データ数を変化させた時のヒストグラムと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。ヒストグラムについては，スコットのルールで最適ビン幅を

表１　各シミュレーションにおける設定

	サンプル数 n	ビン数 m	繰り返し回数 r
ケース①	50, 100, 200, 500, 1000, 5000	$m=Scott$	30
ケース②	50	3, $m=Scott$, 13	2, 5, 10, 20, 50, 100
ケース③	50, 100, 200, 500, 1000, 5000	$m=Scott$	2, 3, 4, …, 20, 50, 100
ケース④	50, 100, 200, 500, 1000, 5000	3, 5, …, 47, 50	30

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，最適なビン数の推定としてスコットのルールを用いる。以降，最適ビン数を用いる時は$m=Scott$と記す。RPHの設定では，mをスコットのルールで算出し，$m-1$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ヒストグラムを構築する。繰り返し上述のヒストグラムを計算し，その平均を取ってRPHを推定し，ISEを計算する。この時，$r=30$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に関して，スコットのルールによるヒストグラム推定を10,000回繰り返し，最適なビン数は平均8.15であった。したがって，最適ビン数約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$m=Scott$，$m=3$，$m=13$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。ここではサンプル数ごとに，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，$m=Scott$で推定する。また，$r=2$，3，4，…，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，$m=3$，5，…, 47, 50とし，各サンプル数及びビン数におけるISE値を計算する。加えて，mを変化させた時の，ISE値を調べる。また，RPHのISE値が，$n=200$, 500におけるヒストグラムのISE値を下回るmの範囲について調べる。

4．シミュレーション結果

4.1．ケース①の結果

表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。表中で，推定精度が良いほどISEは小さいため，比較して値が小さい方に下線を引いてある。ヒストグラムと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。しかしながら，nが大きくなるにつれて両者のISE値の差は小さくなる。

表３は，データ数を変化させた時の，推定値の安全性（頑健性）を表すISE標準偏差の計算結果

表２　ケース①：ISE値（$m=Scott$）

	$n=50$	$n=100$	$n=200$	$n=500$	$n=1000$	$n=5000$
ヒストグラム	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３　ケース①：ISE標準偏差（m＝$Scott$）

	$n=50$	$n=100$	$n=200$	$n=500$	$n=1000$	$n=5000$
ヒストグラム	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

を示す。ただし，表中で，小さい値の方に下線を引いてある。ヒストグラムとRPHのどちらも，nが増加するにつれてISE標準偏差は小さくなる。nに関わらず，RPHの方がISE標準偏差は小さい。しかしながら，nが大きくなるにつれて両者のISE標準偏差の差は小さくなる。ヒストグラムと比較して，RPHの方が安定（頑健）であることが分かる。

4.2．ケース②の結果

表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。表中では，n＝50におけるヒストグラムのISE値0.026160を下回る箇所に下線を引いてある。m＝3，m＝$Scott$，m＝13のどの場合でも，rが増えるにつれてISE値は小さくなる。rに関わらずm＝13の場合，ISE値が相対的に最も小さい。ヒストグラムのISE値と比較して，m＝$Scott$，13の場合では，$r\geq5$でISE値が小さくなる。一方で，m＝3の場合，rを増やしてもヒストグラムより推定精度は劣ることが分かった。したがって，最適ビン数を超えるm＝13で相対的に最も推定精度が優れることから，RPHのビン数はヒストグラムの最適ビン数より多く設定する方が良いと思われる。

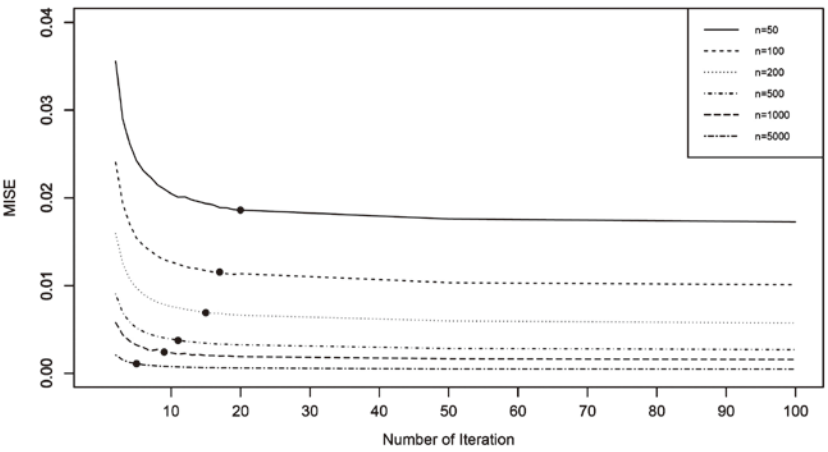
表５は，繰り返し回数を変化させた時のISE標準偏差の計算結果を示す。n＝50におけるヒストグラムのISE標準偏差0.012104を下回る箇所に下線を引いてある。m＝3，m＝$Scott$，m＝13のどの場合も，rが増えるにつれてISE標準偏差は小さくなる。rに関わらずm＝13の場合，ISE標準偏差が相対的に最も小さい。ケース①の結果から，ヒストグラムのISE標準偏差と比較して，m＝$Scott$，m＝13の場合は$r\geq5$，m＝3の場合は$r\geq20$で標準偏差がより小さくなることが分かった。分散については，繰り返し回数に依存するものの，ビン数がm＝$Scott$以上の場合だけでなく，少ない場合であってもヒストグラムより安定化が図られる。

表４　ケース②：ISE値（n=50）

	$r=2$	$r=5$	$r=10$	$r=20$	$r=50$	$r=100$
ヒストグラム	0.026160[3]					
$m=3$	0.097327	0.087727	0.084482	0.083044	0.081971	0.081602
$m=Scott$[4]	0.035597	0.024235	0.020478	0.018602	0.017591	0.017244
$m=13$	0.028028	0.018019	0.014704	0.012856	0.011889	0.011572

表５　ケース②：ISE標準偏差（n=50）

	$r=2$	$r=5$	$r=10$	$r=20$	$r=50$	$r=100$
ヒストグラム	0.012104[3]					
$m=3$	0.026456	0.018560	0.014095	0.010336	0.007370	0.006080
$m=Scott$[4]	0.017645	0.012062	0.010489	0.009986	0.009692	0.009471
$m=13$	0.013654	0.008749	0.007771	0.007420	0.007110	0.007158



(注)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。

図3　各繰り返し回数を変化させたときのサンプル数別のISE値（$m=Scott$）

4.3. ケース③の結果

図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，我々は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る。エルボー法と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。

サンプル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値の減少幅は非常に小さくなっている。このことから，サンプル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，サンプル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。

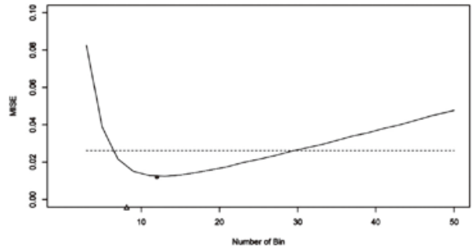
4.4. ケース④の結果

図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によってヒストグラム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くなっている。また，どのnにおいても，RPHのISE値が最小となるmはヒストグラムの最適な$m=Scott$より大きい。

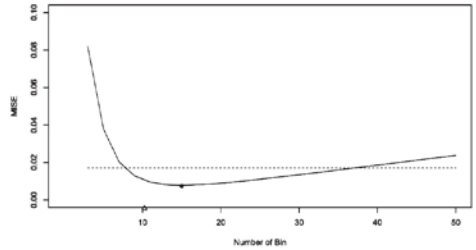
図5は$n=200$のヒストグラムと，$n=100$，200のRPHのISE値を比較したグラフである。グラフ内で，ヒストグラムの$n=200$におけるISE値は0.011116である。ヒストグラムのISE値をRPHが下回っているのは，$n=100$では$10 \leq m \leq 25$，$n=200$では$9 \leq m \leq 48$である。

図6は$n=500$のヒストグラムと，$n=200$，500のRPHのISE値を比較したグラフである。グラフ内で，ヒストグラムの$n=500$におけるISE値は0.006284である。ヒストグラム（$n=500$）のISE値をRPHが下回っているのは，$n=200$では$12 \leq m \leq 27$，$n=500$では$m \geq 11$である。

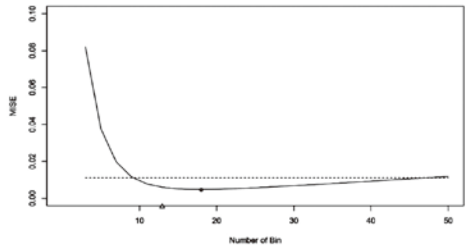
図5，図6から，ヒストグラムより少ないサンプル数でもRPHの推定精度が良い場合があった。そのため，RPHのビン数を適切に選択すればサンプル数が少なくてもヒストグラムよりも優れた推定が可能である。



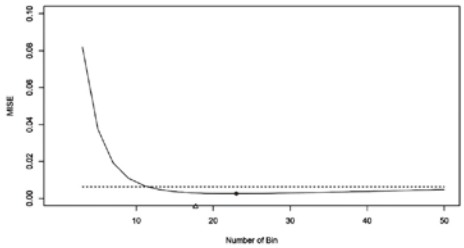
(a)$n = 50$



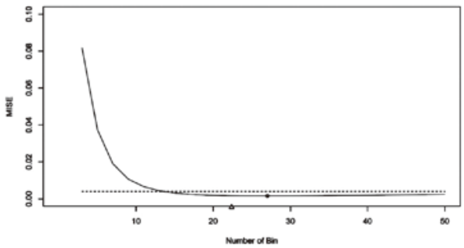
(b)$n = 100$



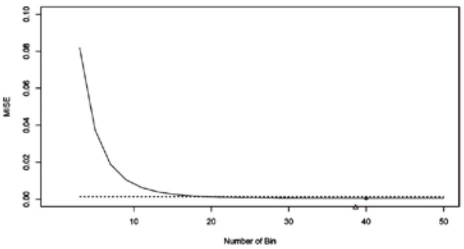
(c)$n = 200$



(d)$n = 500$



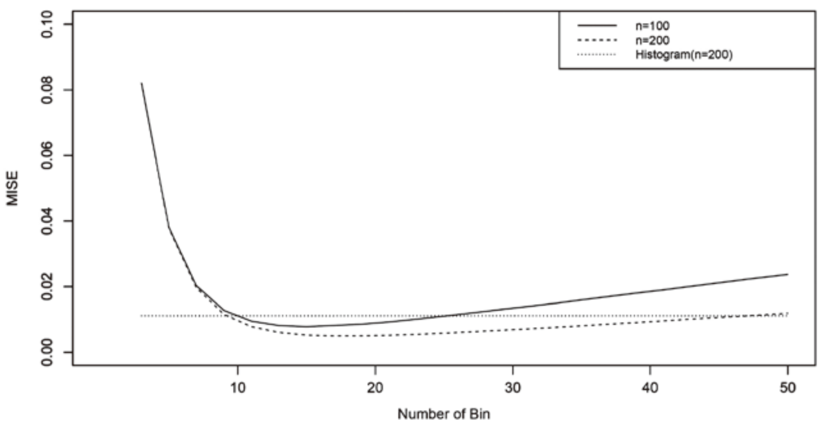
(e)$n = 1000$



(f)$n = 5000$

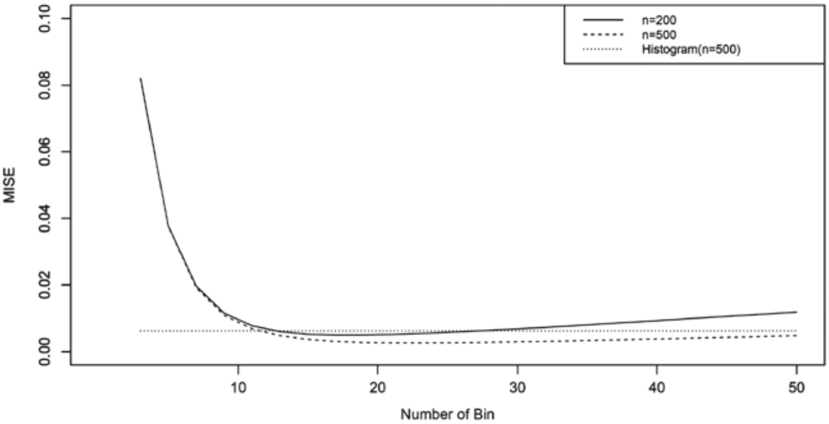
（注）グラフの点線は最適なヒストグラム（$m = Scott$）

図４　サンプル数ごとの各ビン数におけるISE値（r＝30）
（実線：RPH，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）



（注）グラフの点線はn＝200の最適なヒストグラム（$m = Scott$）

図５　ヒストグラム（n＝200）とRPH（r＝30）のISE値の比較



（注）グラフの点線は$n=500$の最適なヒストグラム（$m=Scott$）

図６　ヒストグラム（$n=500$）とRPH（$r=30$）のISE値の比較

表６は，シミュレーション計算で得られたサンプル数ごとのRPHのISE値の最小値及びその時のビン数，ヒストグラム推定シミュレーションにおける平均ビン数，RPHのISE値が最小時のビン数とヒストグラムの平均ビン数の比率である。ただし，表中のmhはヒストグラム推定における平均ビン数とする。ヒストグラムとRPHどちらもm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/mhの平均は1.31であった。加えて上記で示した通り，RPHではヒストグラム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ヒストグラムの1.5倍を推奨する。

５．多峰の分布におけるシミュレーション

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ヒストグラムとRPHの推定精度を比較する。定義域［-6, 6］の混合正規分布$\frac{3}{4}N\left(-1,\left(\frac{3}{2}\right)^2\right)+\frac{1}{4}N\left(2,\left(\frac{1}{3}\right)^2\right)$に従うサンプルについて，ヒストグラムとRPHそれぞれのISE計算シミュレーションを10,000回行い，その結果を比較する。$n=50, 100, 200, 500, 1000, 5000$，$r=30$とする。ヒストグラムのビン数はスコットのルールで推定する。一方で，RPHのmについては4.2節～4.4節の結果から，ヒストグラムより多い方が望ましいことが明らかである。そのため，mはmhを1.5倍した値の整数部分を使用する。表７は各サンプル数におけるヒストグラムの平均

表６　各サンプル数における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	12	15	18	23	27	40
mh	8.15	10.25	12.94	17.70	22.39	38.62
m/mh	1.47	1.46	1.39	1.30	1.21	1.04

表７　ヒストグラムの平均ビン数及び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（$=mh\times1.5$）	10	12	16	21	27	47

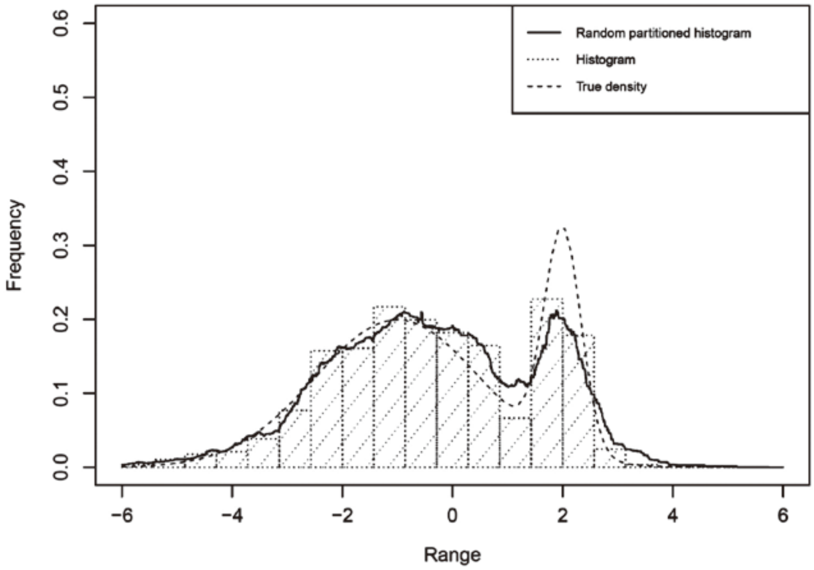


図７　混合正規分布 $\frac{3}{4}N\left(-1,\left(\frac{3}{2}\right)^2\right)+\frac{1}{4}N\left(2,\left(\frac{1}{3}\right)^2\right)$（$n=500$）におけるRPH推定量の例

ビン数と，RPHでの選択ビン数である。また，図７は$n=500$におけるRPH推定量の例である。

まず，各サンプル数におけるヒストグラムとRPHのISE値と標準偏差の結果を示す。

表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。表中で，比較して値が小さい方に下線を引いてある。ヒストグラムと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。しかしながら，nが大きくなるにつれて両者のISE値の差は小さくなる。

表９は，データ数を変化させた時のISE標準偏

表８　ISE値（多峰）

	$n=50$	$n=100$	$n=200$	$n=500$	$n=1000$	$n=5000$
ヒストグラム	0.035698	0.028556	0.016088	0.012439	0.011044	0.003300
RPH（$r=30$）	0.022530	0.017825	0.012525	0.008159	0.005257	0.001620

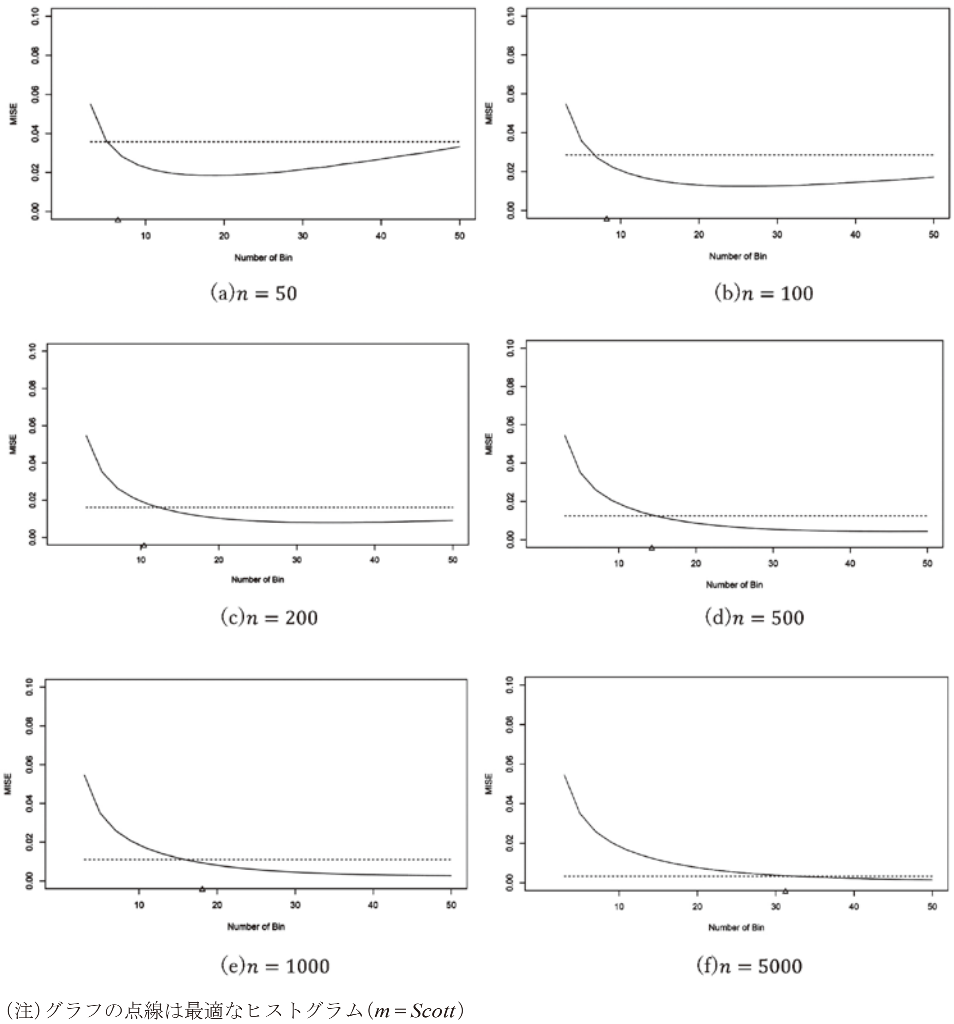
表９　ISE標準偏差（多峰）

	$n=50$	$n=100$	$n=200$	$n=500$	$n=1000$	$n=5000$
ヒストグラム	0.009934	0.005629	0.003052	0.003443	0.001543	0.000188
RPH（$r=30$）	0.004964	0.003822	0.002943	0.001964	0.001360	0.000477

差の計算結果を示す。表中で，比較して値が小さい方に下線を引いてある。ヒストグラムとRPHのどちらも，nが増加するにつれてISE標準偏差は小さくなる。$n=5000$以外では，RPHの方がISE標準偏差は小さい。しかしながら，nが大きくなるにつれて両者のISE標準偏差の差は小さくなる。$n=5000$の場合には，ヒストグラムのビン幅が十分小さく，ビン幅及びそれに伴うビン数がサンプルによって変動しにくいため，ヒストグラムの方がRPHよりも分散が安定化すると考えられる。したがって，小，中サンプルにおいては，RPHの方がヒストグラムよりも分散は安定化している。

図８は各サンプル数で，ビン数を変化させた時のISE値である。nに関わらず，mによってヒストグラム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。

表10はシミュレーション計算で得られたサンプル数ごとのRPHのISE値の最小値及びその時のビン数，ヒストグラム推定シミュレーションにおける平均ビン数，RPHのISE値が最小時のビン数と



(注)グラフの点線は最適なヒストグラム($m=Scott$)

図８　サンプル数ごとの各ビン数におけるISE値（多峰，$r=30$）
（実線：RPH，三角印：ヒストグラムの平均ビン数）

表10　各サンプル数における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	20	24	34	45	57	90
mh	6.48	8.23	10.43	14.28	18.10	31.24
m/mh	3.08	2.92	3.26	3.15	3.15	2.88

ヒストグラムの平均ビン数の比率である。mh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/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ではヒストグラム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。

図９はシミュレーション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ヒストグラムの平均ビン数の比m/mh（表５），および多峰の分布におけるシミュレーションで得られたm/mh（表９）それぞれに対する回帰直線である。横軸はサンプル数で，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/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/mhが約３である。この理由としては，ヒストグラム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のビン数はヒストグラムよりも多い方が望ましく，単峰の場合にはヒストグラムの1.5倍，多峰の場合にはヒストグラムの３倍を目安にビン数選択することを推奨する。

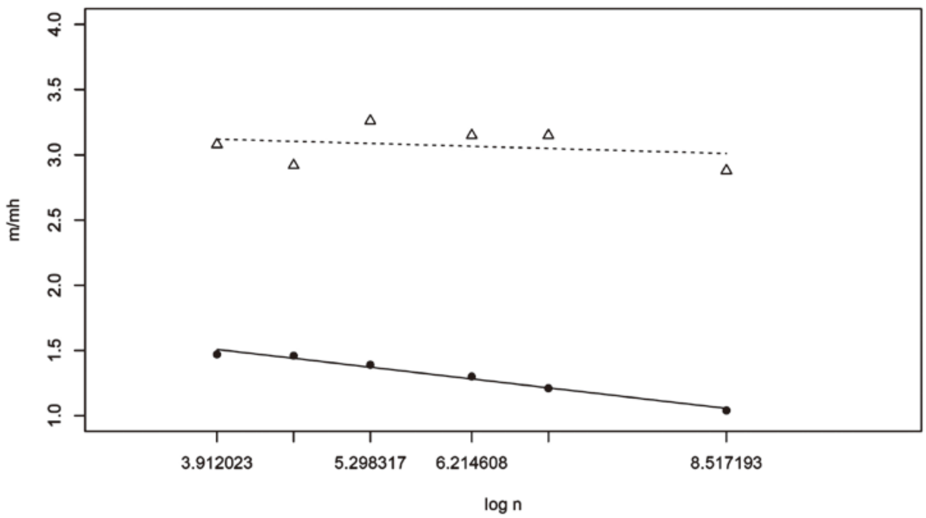


図９　サンプル数ごとのm/mhに対する回帰直線

（横軸：log n，実線：m/mh（単峰）の回帰直線，破線：m/mh（多峰）の回帰直線，
丸印：m/mh（単峰）の実現値，三角印：m/mh（多峰）の実現値）

6．結論と考察

6.1．結論

本稿では，RPHの有限サンプルにおける性質について明らかにする分析を行った。真の分布が単峰の場合について，定義域［-4, 4］の標準正規分布N（0,1）に従うサンプルで4ケースのシミュレーションを行い，それらの結果からヒストグラムとRPHとの推定精度を比較した。

ケース①の結果から，ヒストグラムと比較して，サンプル数に関わらずRPHでISE値の減少と分散の安定化が両方得られることが明らかになった。

ケース②の結果から，ビン数について，ヒストグラムの最適ビン数及びそれを超えて設定したRPHについては，繰り返し回数が極端に少ない時以外ではヒストグラムよりも推定精度が改良された。また，最適ビン数を超えるビン数を設定した場合が相対的に推定精度が良かった。

ケース③の結果から，サンプル数ごとに，各繰り返し回数におけるISE値をプロットするとエルボーカーブと類似したグラフが得られた。このことから，曲線がなだらかになる一番左端の箇所で繰り返し回数を選択することが望ましい。繰り返し回数の目安としては，サンプル数が小さい場合には20以上と比較的多くした方が良い推定が得られる。一方で，サンプル数が大きい場合には繰り返し回数が5回程度であっても優れた推定精度となる。

ケース④の結果から，サンプル数に関わらず，RPHの方がビン数によってヒストグラムより推定精度が良くなることが明らかになった。また，ビン数によってヒストグラムよりも少ないサンプル数であっても推定精度が良くなることが分かった。更に，RPHのシミュレーション計算から探索的に得られた各サンプル数における最小のISE値とその時のビン数を示した。このビン数とヒストグラムの平均ビン数との比率を計算すると平均1.31だった。

真の分布が多峰の場合については，定義域［-6, 6］の混合正規分布$\frac{3}{4}N\left(-1,\left(\frac{3}{2}\right)^2\right)+\frac{1}{4}N\left(2,\left(\frac{1}{3}\right)^2\right)$に従うサンプルを用いて，ヒストグラムとRPHの推定精度を比較した。ISE値と標準偏差の計算結果から，ヒストグラムと比較すると，RPHでISE値の減少と分散の安定化が両方得られることが明らかになった。また，単峰の場合と同様に，サンプル数に関わらず，RPHの方がビン数によってヒストグラムより推定精度は良くなることが明らかになった。更に，シミュレーション結果から探索的に得られたRPHの各サンプル数における最小ISE値とその時のビン数を示した。このビン数とヒストグラムの平均ビン数との比率を計算すると平均3.07だった。以上より，単峰・多峰の分布の形状に関係なくヒストグラムよりも多いビン数を選択する方がRPHの推定精度が良くなることが分かった。

6.2．考察

本稿でのシミュレーションの結果から，RPH推定に必要となるビン数，繰り返し回数を上手く調節することで，ヒストグラムよりも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。

ビン幅の選択に関して，通常のヒストグラムでは最適ビン幅よりも狭いビン幅（最適よりも多いビン数）を選択すると平滑化不足により推定精度が悪くなる。反対に，最適より広いビン幅（最適より少ないビン数）を選択すると平滑化過多でやはり推定精度が悪くなる。RPHでは多少の平滑化過多及び平滑化不足を許してヒストグラムよりも推定精度が改良される。そのため，RPHのビン数選択については，許容できる範囲が広いものと考察される。ただし，ヒストグラムと比較して平滑化不足の場合に推定精度が高い傾向にある。したがって，シミュレーション実験の結果を踏まえ，RPHのビン数については，分布に依存するため目安値ではあるが，実用的には単峰の場合はヒストグラムの1.5倍，多峰の場合はヒストグラムの3倍で選択することを推奨する。

単峰及び多峰のどちらの場合でもヒストグラムと比較して，RPHの方がMISEの意味で推定精

度が改良され，ほとんどの場合で分散が安定化した。このことから，分布の形状に関わらず本手法が有効に働くことが示された。したがって，ビン数や繰り返し回数等のパラメータを適切に選択すれば，様々な分布に適用可能な汎用的手法である。

本稿では，RPHについて有限サンプルでの性質をシミュレーションにより調べ，実用性について議論した。理論的にはビン数ではなく，ビン幅推定の問題であるため，RPHの最適ビン幅も含めた大標本における漸近的性質については今後明らかにしたい。また，本稿では定義域を与えて議論したが，定義域は未知の場合の方が多い。そのため，定義域を任意に設定できる場合には，定義域を広く取ることでデータの集中度が低い箇所でRPH推定量が上方へ押し上げられる傾向は軽減できると考えられる。定義域の広さに関する推定精度への影響や適切な選択法については今後の検討課題である。

【注】

1) サンプル数をn，サンプル標準偏差を$\hat{\sigma}$とすると，スコットのルールにおけるビン幅$\hat{h}$の推定式は，

$$\hat{h} = 3.5\,\hat{\sigma} n^{-1/3},$$

であり，シミュレーションでもビン数をスコットのルールから推定する際にはこの式を用いる。

2) スタージェスのルールとは，Sturges（1926）が提案したヒストグラムのビン数決定法であり，広く普及している手法である。サンプル数をnとすると，ビン数$\hat{m}$の決定式は以下の通りである。

$$\hat{m} = \log_2 n + 1.$$

3) ヒストグラムのISE値と標準偏差値についてはRPHとの比較を目的として表に掲載している。そのため，繰り返し回数rの項目はヒストグラムには関係しない。

4) 10,000回のシミュレーションにおいて，スコットのルールで推定したヒストグラムの平均ビン数は8.15であった。

【引用・参考文献】

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

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

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

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

(5) Jean-Pierre Lecoutre, "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, 1987, pp.265-276.

(6) M. Shibuya and H. Yamato, "Characterization of Some Random Partitions", *Japan Journal of Industrial and Applied Mathematics*, 12, 1995, pp.237-263.