平均に対する平滑化ブートストラップ法におけるバンド幅の選択に関する一考察

難波, 明生

(Citation)
国民経済雑誌,205(3):41-55

(Issue Date)
2012-03

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81008394

(URL)
https://hdl.handle.net/20.500.14094/81008394

平均に対する平滑化ブートストラップ法におけるバンド幅の選択に関する一考察

難　波　明　生

国民経済雑誌　第 205 巻　第 3 号　抜刷
平 成 24 年 3 月

平均に対する平滑化ブートストラップ法におけるバンド幅の選択に関する一考察

難　波　明　生

本稿では，平滑化ブートストラップ法（smoothed bootstrap）を用いて平均の信頼区間を求めることを考える。平滑化ブートストラップ法を用いる際にはバンド幅（bandwidth）というパラメータを設定する必要がある。平滑化ブートストラップ法におけるバンド幅の選択についてはいくつかの研究があるが，本稿では，分析対象である未知の分布が正規分布であると仮定し，密度関数および分布関数の推定量の平均自乗誤差の積分値が最小となるようなバンド幅を利用することを考える。シミュレーションの結果から，分布関数の推定の平均自乗誤差の積分値を最小にするようなバンド幅を利用した場合，平滑化ブートストラップ法による信頼区間は非常に優れた性質を持つことが示される。

キーワード　　ノンパラメトリック法, カーネル密度推定，ブートストラップ法

1　は　じ　め　に

母集団の平均に対する信頼区間を求めたい場合，母集団が正規母集団であれば t 分布を用いて正確な信頼区間を容易に求めることができる。しかしながら，母集団の分布が複雑である場合や未知である場合，正確な信頼区間を求めることは一般に容易ではない。このような場合には，中心極限定理に基づき，正規分布による近似が通常用いられている。しかし，小標本においては正規分布による近似もあまり正確ではない場合が多い。このような場合に有効なのが Efron（1979）により提案されたブートストラップ法である。ブートストラップ法を用いれば，通常の正規近似よりも正確な信頼区間が得られる可能性があることが Beran（1988), Hall（1992）により示されている。

しかしながら，ブートストラップ法にもいくつかの問題がある。その一つが，通常のブートストラップ法では手元にある標本からの無作為抽出を行うので，元々手元にある標本以外の標本値は得られないという点である。「手元にある標本からの無作為抽出を行う」という，通常のブートストラップ法で用いられる手順は，「標本の経験分布から無作為抽出を行う」ということを意味している。また，経験分布は，分布関数のもっとも単純な推定値であると

考えることができる。このことから，経験分布という単純な分布関数の推定値ではなく，分布のより精密な推定値を用いて，その分布からの無作為抽出を行えば良いのではないかという発想が生じてくるのは極めて自然なことである。このような方法は，経験分布が不連続な階段関数であるのに対し，分布関数の推定値が滑らかな関数になることから，平滑化ブートストラップ法（smoothed bootstrap）と呼ばれている。

上記のように，平滑化ブートストラップ法においては，まず母集団の分布を推定する必要がある。母集団の分布が，例えば正規分布であるというように，ある特定の分布であると分かっている場合には，未知パラメータを最尤法などの方法で推定することで分布の推定を行うことができる。しかし，実際の問題では，多くの場合母集団の分布が未知である。このような場合に用いられるのが，分布の形状に特定の仮定を置かず，密度関数を推定することができるカーネル密度推定（kernel density estimate）と呼ばれる方法である。

カーネル密度推定を用いる際には，分析者はカーネル関数と呼ばれる関数とバンド幅（bandwidth）の値を定めなければならない。一般に，カーネル関数は推定結果に大きな影響を与えないが，バンド幅の選択は非常に大きな影響を及ぼすことが知られている。したがって，バンド幅（bandwidth）の選択は非常に重要であると考えられる。多くの場合，バンド幅はカーネル密度推定量の MISE（mean integrated squared error）を最小にするように選ぶことが望ましいと考えられているが，MISE は推定の対象である未知の確率密度関数に依存した値となる。したがって，推定の対象となっている確率密度関数が分からなければ，MISE を最小化するような最適なバンド幅の値を求めることはできない。また，カーネル密度推定量を積分することで分布関数の推定量が得られるが，MISE を最小にするようなバンド幅は，我々が確率密度関数を正確に推定したいのか，それとも分布関数を推定したいのかによって異なる値になることが知られている。さらに，分布関数の推定量の MISE を最小化するバンド幅もまた未知パラメータに依存する。したがって，MISE を最小化するようなバンド幅を用いる場合には，密度関数と分布関数のどちらの推定量の MISE を最小化するにしても，未知の確率密度関数に何からの関数を仮定して，未知パラメータを推定量で置き換える等の方法を用いる必要がある。

Silverman and Young（1987），Hall, DiCiccio and Romano（1989），De Angelis and Young（1992），Polansky and Schucany（1997），El-Nouty and Guillou（2000）等はバンド幅を適切に選んだ時，平滑化ブートストラップ法を用いることにより，通常のブートストラップ法よりも推定の精度を高めることができる場合があることを示した。しかし，このようなバンド幅は一般に未知パラメータに依存し，未知パラメータを推定値で置き換えた場合には平滑化ブートストラップ法の精度はあまり高くないことが Polansky and Schucany（1997）のシミュレーションにより示されている。また，平滑化ブートストラップ法の精度を高めることのできる

バンド幅は，MISE を最小化することにより得られるバンド幅とは一般に全く異なるものである。

しかしながら，カーネル密度推定量の MISE を最小化するということは，カーネル密度推定により推定された確率密度関数が，全体として真の確率密度関数に近いことを意味する。このことを考えれば，MISE を最小化することにより得られるバンド幅を用いて平滑化ブートストラップ法を行えば，正規分布による近似や，通常のブートストラップ法を用いた場合よりも良い信頼区間が得られる可能性があるのではないかと思われる。さらに，通常のブートストラップ法は分布関数を経験分布関数で推定したものと解釈できることを考えると，平滑化ブートストラップ法においても，確率密度関数の推定量の MISE を最小化させるようにバンド幅を選択するよりも，分布関数の推定量の MISE を最小化させるようなバンド幅を用いた方が良いのではないかという発想が浮かぶのも自然なことであろう。

したがって，本稿では，いくつかの分布の平均について，真の分布が正規分布であると仮定して MISE を最小化することによって得られるバンド幅を用いた場合，平滑化ブートストラップ法による信頼区間がどのような特性を持つのかをシミュレーションにより分析する。前述のように，確率密度関数の推定量を考える場合と，分布関数を推定量を考える場合によって，MISE を最小化するバンド幅は異なった値を取る。さらに，いずれのバンド幅も未知の確率密度関数に依存するので，未知の分布は正規分布であると仮定してバンド幅を推定し，得られたバンド幅を用いて，シミュレーションを行う。本稿の構成は以下の通りである。第 2 節では，平滑化ブートストラップ法を用いる準備として，カーネル密度推定法について説明する。第 3 節ではバンド幅の選択方法および MISE について説明する。第 4 節ではブートストラップ法を，通常のブートストラップ法と平滑化ブートストラップ法の違いを述べながら説明する。第 5 節でシミュレーションにより，平滑化ブートストラップ法によって得られる信頼区間の精度を分析する。

2　カーネル密度推定

$X_1, X_2, \ldots, X_n$ は確率密度関数 $f(x)$ を持つ確率分布から独立に得られた標本であるとする。$f(x)$ の関数型が，例えば平均 μ，分散 σ^2 の正規分布の確率密度関数

$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad (1)$$

であることが既知であれば，最尤法などの推定法を用いてパラメータ μ, σ^2 を推定することにより，$f(x)$ の推定量を得ることができる。このような推定法は，パラメトリック推定と呼ばれる。パラメトリックな推定法を行う場合は，上の例のように $f(x)$ の関数型を前もって定式化しなければならない。しかし，$f(x)$ の関数型は多くの場合未知であるため，定式

化の誤りが生じる可能性がある。

これに対し，ノンパラメトリック推定では，$f(x)$ の関数型を前もって定めることなく推定を行うので，定式化の誤りが生じる可能性は無い。確率密度関数をノンパラメトリックに推定する最も簡単な方法は次のようなものである。

$F(x)$ を $f(x)$ に対応する分布関数とすると，

$$\begin{aligned} F(x) &= \mathrm{P}(X \leqq x) \\ &= \int_{-\infty}^{x} f(x)dx \end{aligned} \tag{2}$$

であるから

$$\begin{aligned} f(x) &= \frac{d}{dx}F(x) \\ &\approx \frac{F(x+h/2)-F(x-h/2)}{h} \\ &= \frac{\mathrm{P}(x-h/2 \leqq X \leqq x+h/2)}{h} \\ &= \frac{P\left(-\frac{1}{2} \leqq \frac{X-x}{h} \leqq \frac{1}{2}\right)}{h} \end{aligned} \tag{3}$$

となる。(3)の分子を推定値で置き換えることにより，$f(x)$ は

$$\begin{aligned} \hat{f}_n(x) &= \frac{\left(\left[-\frac{1}{2}, \frac{1}{2}\right] \text{に入っている} \frac{X_i-x}{h} \text{の個数}\right)/n}{h} \\ &= \frac{1}{nh}\sum_{i=1}^{n} I\left(-\frac{1}{2} \leqq \frac{X-x}{h} \leqq \frac{1}{2}\right) \end{aligned} \tag{4}$$

のように推定される。ただし，$I(A)$ は事象 A が起こったときに 1，それ以外では 0 の値を取る関数であり，indicator function と呼ばれる。また，$h(>0)$ はバンド幅と呼ばれる，分析者が定めるパラメータである。この推定量は naive estimator, local histogram estimator 等と呼ばれている。

(4)で与えられる推定量は，local histogram estimator という名前も示しているように，滑らかな関数ではない。そこで，Rosenblatt (1956) は(4)を拡張し，次のような推定量を考案した。

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) \tag{5}$$

h は(4)の場合同様にバンド幅である。また，$K(\cdot)$ はカーネルと呼ばれる関数である。容易に分かるように，$\hat{f}(x)$ は x について滑らかな関数である。$\hat{f}(x)$ が一致性を持つために，

通常以下の仮定が置かれる。ただし，表記を簡単にするために，今後は積分範囲が明示されていない場合には，積分範囲は積分する変数の定義域全体であるとする。

仮定 1　カーネル $K(\cdot)$ は以下の性質を満たすものとする。

（ⅰ）．$\int K(v)dv=1$

（ⅱ）．$K(v)=K(-v)$

（ⅲ）．$\int v^2K(v)dv=\kappa_2>0$

ここで，仮定 1 の内容を簡単に見ておこう。（ⅰ）が満たされることにより，

$$\begin{aligned}\int \hat{f}(x)dx&=\int \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)dx\\&=\frac{1}{nh}\sum_{i=1}^{n}\int K\left(\frac{X_i-x}{h}\right)dx\\&=\frac{1}{nh}\sum_{i=1}^{n}\int K(\phi_i)hd\phi_i\\&=\frac{1}{n}\sum_{i=1}^{n}\int K(\phi_i)d\phi_i=1\end{aligned}\qquad(6)$$

となる。ただし，$\phi_i=(X_i-x)/h$ である。このことから，$K(\cdot)\geqq 0$ であれば，$\hat{f}(x)\geqq 0$ かつ $\int \hat{f}(x)dx=1$ となるので，カーネル密度推定量 $\hat{f}(x)$ は確率密度関数であるといえる。また，（ⅱ）は $K(v)$ が $v=0$ について左右対称であることを意味する。（ⅰ），（ⅱ）を満たす $K(\cdot)$ としては $v=0$ について左右対称である確率密度関数を用いれば良い。$K(\cdot)$ として（ⅰ），（ⅱ）を満たすような確率密度関数を用いた場合，（ⅲ）は確率密度関数 $K(\cdot)$ を持つ分布の分散が有限な正の値を持てば良いことを意味している。

$X_i, X_2, \ldots, X_n$ が 3 回微分可能な確率密度関数 $f(x)$ を持つ確率分布から独立に得られた標本であるとすると，仮定 1 のもとで $\hat{f}(x)$ のバイアスは次のように計算される。

$$\begin{aligned}\mathrm{Bias}[\hat{f}(x)]&=\mathrm{E}[\hat{f}(x)]-f(x)\\&=\mathrm{E}\left[\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)\right]-f(x)\\&=h^{-1}\mathrm{E}\left[K\left(\frac{X_1-x}{h}\right)\right]-f(x)\\&=h^{-1}\int f(x_1)K\left(\frac{x_1-x}{h}\right)dx_1-f(x)\\&=h^{-1}\int f(x+hv)K(v)hdv-f(x)\end{aligned}$$

$$=\int\left\{f(x)+f^{(1)}(x)hv+\frac{1}{2}f^{(2)}(x)h^2v^2+O(h^3)\right\}k(v)dv-f(x)$$

$$=\frac{h^2}{2}f^{(2)}(x)\kappa_2+O(h^3) \qquad (7)$$

2行目から3行目の変形では X_i が独立に同一の分布に従うこと，4行目から5行目の変形では $(x_1-x)/h=v$ という変数変換，5行目から6行目への変形では $f(x+hv)$ のテーラー展開を用いている。

同様の計算により，$\hat{f}(x)$ の分散は

$$\mathrm{Var}[\hat{f}(x)]=\mathrm{E}[\hat{f}(x)-\mathrm{E}[\hat{f}(x)]]$$

$$=\frac{1}{nh}\{\kappa f(x)+O(h)\} \qquad (8)$$

であることを示すことができる。ただし，$\kappa=\int K^2(v)dv$ である。(7)と(8)を用いれば，$\hat{f}(x)$ の平均自乗誤差（mean squared error; MSE）は

$$\mathrm{MSE}[\hat{f}(x)=\mathrm{E}[(\hat{f}(x)-f(x))^2]$$

$$=\mathrm{Var}[\hat{f}(x)]+(\mathrm{Bias}[\hat{f}(x)])^2$$

$$=\frac{h^4}{4}[\kappa_2 f^{(2)}(x)]^2+\frac{\kappa f(x)}{nh}+o(h^4+(nh)^{-1})=O(h^4+(nh)^{-1}) \qquad (9)$$

となる。したがって，$c>0, \beta>1$ を定数とし $h=cn^{-1/\beta}$ のようにバンド幅を選択すれば，$n\to\infty$ の時 $h\to 0, nh\to\infty$ となり $\hat{f}(x)$ は $f(x)$ に平均自乗収束する。したがって，バンド幅が以上の条件を満たせば，$\hat{f}(x)$ は $f(x)$ の一致推定量であるといえる。

さらに，上でみたように $\int\hat{f}(x)dx=1$ であるから，$\hat{f}(x)$ を x について積分したものは分布関数であると考えられるので，$\hat{f}(x)$ を $X_i(i=1, 2, \ldots, n)$ の分布関数 $F(x)$ の推定量として

$$\hat{F}(x)=\int_{-\infty}^{x}\hat{f}(x)dz=\frac{1}{n}\sum_{i=1}^{n}G\left(\frac{x-X_i}{h}\right) \qquad (10)$$

として用いることは自然であろう。ただし $G(x)=\int_{-\infty}^{x}K(x)dx$ である。特定の条件の下で，$\hat{f}(x)$ の場合と同様の計算により

$$\mathrm{MSE}[\hat{F}(x)]=\mathrm{E}[(\hat{F}(x)-F(x))^2] \qquad (11)$$

$$=a_0(x)n^{-1}-a_1(x)hn^{-1}+a_2(x)h^4+o(h^4+hn^{-1}) \qquad (12)$$

であることが示される。ただし，$a_0(x)=F(x)(1-F(x))$, $a_1(x)=\alpha_0 f(x)$, $\alpha_0=2\int vG(v)\times K(v)dv$, $a_2(x)=[(\kappa/2)f'(x)]^2$ であり，h はある $0<\varepsilon<1/8$ に対して $0\leqq h\leqq Cn^{-\varepsilon}$ を満

たすものとする（証明は Li and Racine［2007］参照）。h がこの条件を満たすとき，明らかに $n \to \infty$ の時 $h \to 0, hn^{-1} \to 0$ であり，$\hat{F}(x)$ は $F(x)$ に平均自乗収束するので，$\hat{F}(x)$ も $F(x)$ の一致推定量である。

上記のように，カーネル密度推定を用いれば，確率密度関数および分布関数の一致推定量を得ることができる。しかしカーネル密度推定量の定義からも明らかなように，カーネル密度推定を用いるためには，分析者はカーネル $K(\cdot)$ およびバンド幅 h を選択しなければならない。一般にカーネルの関数型が結果に及ぼす影響はさほど大きくないが，バンド幅が及ぼす影響は非常に大きいといわれている。したがって，以下ではバンド幅の選択方法について説明する。

2.1　Mean Integrated Squared Error とバンド幅の選択

カーネル密度推定におけるバンド幅の選択基準の一つは，（9）で与えられる MSE を x について積分した mean integrated squared error（MISE）

$$\begin{aligned}\mathrm{MISE}[\hat{f}(x)] &= \int \mathrm{MSE}[\hat{f}(x)]dx \\ &= \int \left\{ \frac{h^4}{4}[\kappa_2 f^{(2)}(x)]^2 + \frac{\kappa f(x)}{nh} + o(h^4 + (nh)^{-1}) \right\} dx \\ &= \frac{1}{4}h^4\kappa_2^2 \int [f^{(2)}(x)]^2 dx + \frac{\kappa}{nh} + o(h^4 + (nh)^{-1}) \qquad (13)\end{aligned}$$

を最小にすることである。そのようなバンド幅は(13)を h に関して微分して 0 とおくことで

$$h_1 = c_0 n^{-1/5} \qquad (14)$$

であることが容易に分かる。ただし

$$c_0 = \kappa^{1/5}\kappa_2^{-2/5}\left\{ \int [f^{(2)}(x)]^2 dx \right\}^{-1/5} \qquad (15)$$

は正の定数である。したがって，h_1 は $f^{(2)}(x)$ に依存することが分かるが，$f^{(2)}$ は通常未知であるため，実際には h_1 をバンド幅として利用することはできない。このバンド幅を利用する方法の一つは，$f(x)$ に何らかの特定の分布（例えば正規分布）を仮定して(15)の値を計算し，必要に応じて $f(x)$ に含まれるパラメータを推定値で置き換えて利用することである。例えば，カーネル $K(\cdot)$ として標準正規分布の確率密度関数を用いた場合，$f(x)$ が平均 μ，分散 σ^2 の正規分布の確率密度関数であると仮定すれば，

$$h_1 = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5} \approx 1.06\sigma n^{-1/5} \qquad (16)$$

となる。したがって，分析者は h として $\hat{h}_1 = 1.06\hat{\sigma}n^{-1/5}$ を用いることができる。ただし，$\hat{\sigma}$

は $X_i, X_2, \ldots, X_n$ から得られる標本標準偏差である。このようなバンド幅の選択方法をプラグ・イン（plug-in）と呼ぶ。この方法の問題点は，$f(x)$ に何らかの分布を仮定するため，この仮定が真の分布からかけ離れている場合には，推定量の精度がかなり落ちる可能性があることである。

確率密度関数の推定の場合と同様に，分布関数の推定の場合にも(12)を積分することで $\hat{F}(x)$ の MISE

$$\begin{aligned}\mathrm{MISE}[\hat{F}(x)] &= \int \mathrm{MSE}[\hat{F}(x)]dx \\ &= A_0 n^{-1} - A_1 h n^{-1} + A_2 h^4 + o(h^4 + hn^{-1}) \end{aligned} \tag{17}$$

が得られる。ただし $A_j = \int a_j(x)dx$ $(j=0, 1, 2)$ である。(17)を h について微分して 0 とおくことで，分布関数の推定量の MISE を最小化するバンド幅は

$$h_2 = \left(\frac{A_1}{4A_2}\right)^{1/3} n^{-1/3} \tag{18}$$

であることがわかる。A_1, A_2 は未知の確率密度関数 $f(x)$ に依存するので，h_1 同様，h_2 もこのままの形でバンド幅として用いることはできない。したがって，この場合にもプラグ・イン等の方法を用いて，実行可能なバンド幅を求めなければならない。密度関数の推定の場合と同様に，カーネル $K(\cdot)$ として標準正規分布の確率密度関数を用いた場合には，$f(x)$ が平均 μ，分散 σ^2 の正規分布であると仮定すれば

$$h_2 = 4^{1/3}\sigma n^{-1/3} \approx 1.59\sigma n^{-1/3} \tag{19}$$

となる。よって，この場合には $\hat{h}_2 = 1.59\hat{\sigma}n^{-1/3}$ をバンド幅として用いることができる。

上記のように，確率密度関数の推定量の MISE を最小化する場合と，分布関数の推定量の MISE を最小化する場合では，得られるバンド幅のオーダーはそれぞれ $n^{-1/5}$ と $n^{-1/3}$ と全く異なったものになっている点に注意が必要である。

3　ブートストラップ法

ブートストラップ法は，手元にある標本を用いて新たな標本を発生させ，得られた標本を用いて計算された統計量により，元の標本から得られる統計量の分布を近似する方法である。手元にある標本を用いて新たな標本を発生させる方法を，リサンプリング（resampling）と呼ぶ。ブートストラップ法を用いることにより，統計量の漸近分布しか分かっていない場合でも，小標本分布をより正確に近似できる場合があることが知られている。以下では，通常のブートストラップ法と平滑化ブートストラップ法について説明する。

3.1　通常のブートストラップ法

通常のブートストラップ法は以下のような手順で行われる。

1．手元にある標本 $X_1, X_2, \ldots, X_n$ から重複を許して大きさ n の標本を無作為抽出する。抽出した標本を $X_1^*, X_2^*, \ldots, X_n^*$ とし，ブートストラップ標本と呼ぶ。

2．ステップ 1 で得られた標本 $X_1^*, X_2^*, \ldots, X_n^*$ 用いて，分布を近似したい統計量の値を計算する。ただし，未知パラメータは $X_1, X_2, \ldots, X_n$ を用いて計算される推定値で置き換える。本稿では，分布の平均に対する信頼区間を求めたいので，

$$\overline{X}-\mu \tag{20}$$

および t 統計量

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \tag{21}$$

の分布を近似することを考える。ただし，μ, σ は $X_i(i=1, \ldots n)$ の平均および標準偏差，$\overline{X}$ は $X_1, X_2, \ldots, X_n$ の標本平均である。上記の 2 つの統計量のいずれかの分布が分かれば，μ の信頼区間を求めることができる。したがって，

$$\overline{X}^*-\overline{X} \tag{22}$$

および

$$\frac{\overline{X}^*-\overline{X}}{s/\sqrt{n}} \tag{23}$$

を計算すれば良いことになる。ただし，$\overline{X}^*$ は $X_1^*, X_2^*, \ldots, X_n^*$ の標本平均，s は $X_1, X_2, \ldots, X_n$ の標本標準偏差である。

3．ステップ 1-2 を B 回繰り返し，B 個の(22)，(23)の推定値を得る。得られた B 個の推定値の経験分布を用いて統計量の分布を近似する。

上記のステップ 2 の(23)のように，t 統計量をブートストラップ法で近似する方法はパーセンタイル t 法と呼ばれている。これに対し，(22)を用いて分布を近似する方法はパーセンタイル法と呼ばれている。パーセンタイル t 法により近似される t 統計量の漸近分布は標準正規分布なので未知パラメータを含まない。これに対し，パーセンタイル法で用いられる，(20)で与えられる統計量の漸近分布は未知パラメータ σ に依存する。通常のブートストラップ法を用いる場合，パーセンタイル t 法を用いて未知パラメータを含まない漸近分布を近似することにより，パーセンタイル法を用いるよりも信頼区間の精度を上げることができることが Beran (1988), Hall (1992) 等により示されている。ブートストラップ法を用いない場合には，中心極限定理に基づき，通常(21)の分布を標準正規分布で近似することにより，μ の信頼区間を求めることになる。

以上のように，通常のブートストラップ法では $\chi=\{X_1, X_2, \ldots, X_n\}$ を所与として，$X_1, X_2, \ldots, X_n$ の経験分布，つまり $\mathrm{P}(X=X_i)=\frac{1}{n}$ である分布から得られた $X_1^*, X_2^*, \ldots, X_n^*$ を用いて統計量の分布を近似していることになる。つまり，通常のブートストラップ法には，(1)ブートストラップ標本が $X_1, X_2, \ldots, X_n$ 以外の値は取らない（ブートストラップ標本を抽出する分布の分布関数が不連続である），(2)ブートストラップ標本として $X_1, X_2, \ldots, X_n$ が選ばれる確率が真の $f(x)$ に関係なく一定である，等の問題点がある。

3.2　平滑化ブートストラップ法

上記のような通常のブートストラップ法の問題点を避けることができるのが，平滑化ブートストラップ法である。通常のブートストラップ法ではブートストラップ標本を $X_1, X_2, \ldots, X_n$ の経験分布から抽出するのに対し，平滑化ブートストラップ法では，ブートストラップ標本は $X_1, X_2, \ldots, X_n$ にカーネル密度推定を応用して得られた分布から抽出する。したがって，平滑化ブートストラップ法は以下のような手順で行われることになる。

1．$X_1, X_2, \ldots, X_n$ を用いてカーネル密度推定量 $\hat{f}(x)$ を求める。

2．ステップ 1 で求めた $\hat{f}(x)$ を確率密度関数とする分布から大きさ n の標本を無作為抽出する。抽出した標本を $X_1^s, X_2^s, \ldots, X_n^s$ とする。

3．ステップ 2 で得られた標本 $X_1^s, X_2^s, \ldots, X_n^s$ を用いて，分布を近似したい統計量の値を計算する。通常のブートストラップ法と同様に未知パラメータは $X_1, X_2, \ldots, X_n$ を用いて計算される推定値で置き換える。ここでは，分布の平均に対する信頼区間を求めたいので，通常のブートストラップ法の場合と同様に，

$$\overline{X}-\mu \tag{24}$$

および

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \tag{25}$$

の分布を近似することを考え，

$$\overline{X}^s-\overline{X} \tag{26}$$

および

$$\frac{\overline{X}^s-\overline{X}}{s/\sqrt{n}} \tag{27}$$

を計算する。ただし，$\overline{X}^s$ は $X_1^s, X_2^s, \ldots, X_n^s$ の標本平均である。

4．ステップ 2-3 を B 回繰り返し，B 個の(26)，(27)の推定値を得る。得られた B 個の推定値の経験分布を用いて統計量の分布を近似する。

平滑化ブートストラップ法では，パーセンタイル t 法を用いなくても，バンド幅の選択によって信頼区間の精度を向上させることができることが Polansky and Schucany (1997), El-Nouty and Guillou (2000) 等により示されている。

ステップ 1 におけるカーネル密度推定を行うためには，バンド幅を選択しなければならない。Polansky and Schucany (1997) は平滑化ブートストラップ法をパーセンタイル法に用いた場合の信頼区間の精度を上げるために最適なバンド幅を導出したが，求められたバンド幅は未知パラメータに依存する。未知パラメータを推定値で置き換えた場合，得られた信頼区間の精度はあまり高くないことが Polansky and Schucany (1997) のシミュレーションにより示されている。本来，MISE を最小化するバンド幅と，信頼区間の精度を上げるために最適なバンド幅は全く異なるものである。しかし，MISE が小さいということは，全体として $\hat{f}(x)$ の推定誤差が小さいことを意味する。そこで，本稿では，前節で説明したプラグ・インと呼ばれる手法を用いてバンド幅を選択する。

また，ステップ 2 では $\hat{f}(x)$ を確率密度関数として持つ分布からの標本抽出を行う必要があるが，これは実際には以下のような方法で行うことができる (Silverman (1986) を参照)。

1．$X_1, X_2, \ldots, X_n$ から X_i^* を無作為抽出する。

2．カーネル密度推定で用いたカーネル $K(\cdot)$ を確率密度関数とする分布から標本 ε_i を抽出する。$X_i^s = X_i^* + h\varepsilon_i$ とすれば X_i^s は $\hat{f}(x)$ を確率密度関数として持つ分布からの無作為標本となる。

つまり X_i^s は通常のブートストラップ法におけるブートストラップ標本 $X_1^*, X_2^*, \ldots, X_n^*$ に $h\varepsilon_i$ を加えたものである。

4　シミュレーション分析

本節では，前節で説明した信頼区間の特性を，シミュレーションにより比較し，平滑化ブートストラップ法の効果を検証する。

シミュレーションは以下のような設定で行った。まず，比較する信頼区間は，前節で説明した(21)で与えられる t 統計量を正規分布で近似して得られる区間，(20)および(21)の統計量に通常のブートストラップ法と平滑化ブートストラップ法を応用して得られる区間である。

平滑化ブートストラップ法を用いる際に必要となるカーネル $K(\cdot)$ は標準正規分布の確率密度関数を用いる。バンド幅については，第 2 節で説明したように，未知の分布が正規分布であると仮定した場合に MISE を最小化する値にプラグ・インすることで得られる $\hat{h}_1 = 1.06\hat{\sigma}n^{-1/5}$ および $\hat{h}_2 = 1.59\hat{\sigma}n^{-1/3}$ を用いる。前節の説明からも分かるように，$h=0$ とした場合には，平滑化ブートストラップ法は通常のブートストラップ法と同一になる。

信頼区間を求める分布は正規分布 (N), 一様分布 (U), 自由度 2 のカイ 2 乗分布 (X), 自由

度 3 の t 分布（T）を用いた。ただし，全ての分布は平均 0，分散 1 となるように変換を行っている。これらの分布から大きさ 10, 20, 30, 50, 100, 200, 500, 1000 の標本を抽出し，10000 回の繰り返し実験を行い，90%，95%，99% の信頼区間が真の平均 0 を含んでいる比率を計算した。ブートストラップ法における繰り返しの回数は $B=1000$ とした。

シミュレーションの結果は表 1，2 の通りである。表では $n=50$ と $n=500$ の結果を示してあるが，他の場合の結果も同様である。

まず，$n=50$ の場合を見ていこう。標本が正規分布に従う場合は，どの方法を用いても良好な結果が得られている。通常のブートストラップ法（$h=0$ の場合）とパーセンタイル法を用いた信頼区間は若干正確ではないが，パーセンタイル t 法を用いることで，精度が改善していることが分かる。また，$\hat{h}_1, \hat{h}_2$ およびパーセンタイル法，パーセンタイル t 法のいずれを用いた場合も，平滑化ブートストラップ法による信頼区間の精度はかなり良好であるといえる。

標本が一様分布に従う場合は，正規分布の場合と同様の傾向が見られる。つまり，通常のブートストラップ法とパーセンタイル法を用いた信頼区間は若干正確ではないが，パーセンタイル t 法を用いることにより，信頼区間の精度を改善することができる。また，平滑化ブートストラップ法による信頼区間はいずれも良好な性質を持っているようである。

標本がカイ 2 乗分布に従う場合は，どの方法を用いても，信頼区間が真の値を含む比率は若干低い傾向が見られる。しかしながら，平滑化ブートストラップ法による信頼区間は比較的良好な結果を示している。また，標本が大きくなると，どの信頼区間も良好な性質を持つようである。

標本が t 分布に従う場合は，通常のブートストラップ法とパーセンタイル t 法によって得られた信頼区間は非常に低い精度しか示していない。このような結果が得られたのは，パーセンタイル t 法がパーセンタイル法よりも有効であるためにはかなり高次のモーメントが存在しなければならないにもかかわらず，t 分布では自由度より小さい次数のモーメントしか存在しないためではないかと思われる。これに対し，平滑化ブートストラップ法とパーセンタイル法を用いた信頼区間は，比較的正確であるといえるだろう。また，この場合には，通常のブートストラップ法の場合と同様，平滑化ブートストラップ法を用いた場合でも，パーセンタイル t 法による信頼区間はあまり正確ではない。

$n=500$ の場合には，どの方法の場合にも，$n=50$ 場合に比べて信頼区間の精度は改善しており，かなり正確な信頼区間が得られているといえる。また，全体としてみると，平滑化ブートストラップ法を用いた場合には，パーセンタイル t 法とパーセンタイル法の精度にはあまり大きな差がないように見える。また，$\hat{h}_1$ をバンド幅とし，パーセンタイル法を用いた場合の信頼区間は，他の方法で得られた信頼区間および名目値より若干高い比率で真の値

表1　シミュレーションにより得られた信頼係数（$n=50$ の場合）

			Percentile			Percentile-t		
		normal	$h=0$	$h=\hat{h}_1$	$h=\hat{h}_2$	$h=0$	$h=\hat{h}_1$	$h=\hat{h}_2$
N	90%	0.8949	0.8749	0.9111	0.9044	0.8868	0.8886	0.8882
	95%	0.9428	0.9306	0.9554	0.9514	0.9426	0.9424	0.9421
	99%	0.9860	0.9791	0.9899	0.9879	0.9854	0.9869	0.9859
U	90%	0.8901	0.8708	0.9032	0.8982	0.8908	0.8901	0.8902
	95%	0.9374	0.9205	0.9479	0.9434	0.9435	0.9409	0.9412
	99%	0.9840	0.9755	0.9869	0.9847	0.9886	0.9874	0.9878
X	90%	0.8796	0.8563	0.8948	0.8896	0.8817	0.8815	0.8812
	95%	0.9290	0.9093	0.9362	0.9315	0.9380	0.9363	0.9361
	99%	0.9730	0.9585	0.9739	0.9711	0.9858	0.9829	0.9823
T	90%	0.9016	0.8872	0.9227	0.9171	0.8617	0.8711	0.8686
	95%	0.9499	0.9452	0.9651	0.9601	0.9193	0.9290	0.9269
	99%	0.9915	0.9872	0.9934	0.9925	0.9762	0.9821	0.9814

表2　シミュレーションにより得られた信頼係数（$n=500$ の場合）

			Percentile			Percentile-t		
		normal	$h=0$	$h=\hat{h}_1$	$h=\hat{h}_2$	$h=0$	$h=\hat{h}_1$	$h=\hat{h}_2$
N	90%	0.8999	0.8980	0.9119	0.9049	0.8995	0.8992	0.9001
	95%	0.9479	0.9458	0.9549	0.9490	0.9465	0.9456	0.9467
	99%	0.9905	0.9877	0.9918	0.9897	0.9885	0.9889	0.9891
U	90%	0.8979	0.8957	0.9106	0.9020	0.8976	0.8980	0.8986
	95%	0.9525	0.9482	0.9585	0.9535	0.9504	0.9507	0.9508
	99%	0.9901	0.9890	0.9919	0.9904	0.9899	0.9894	0.9898
X	90%	0.8994	0.8965	0.9102	0.9028	0.8989	0.8984	0.8990
	95%	0.9458	0.9426	0.9525	0.9464	0.9473	0.9471	0.9472
	99%	0.9873	0.9838	0.9890	0.9854	0.9887	0.9881	0.9886
T	90%	0.9032	0.9036	0.9162	0.9100	0.8914	0.8922	0.8910
	95%	0.9501	0.9508	0.9611	0.9562	0.9388	0.9415	0.9399
	99%	0.9914	0.9921	0.9940	0.9925	0.9847	0.9857	0.9859

を含むようであるが，$\hat{h}_2$ をバンド幅として用い，パーセンタイル法により求めた信頼区間は，どの場合にもかなり良好な性質を示している。

以上のシミュレーション結果から，$\hat{h}_2$ を用いて平滑化ブートストラップ法を行い，パーセンタイル法により信頼区間を求めるという方法は，非常に有効な方法であると考えられる。

パーセンタイル t 法により，パーセンタイル法の精度を改善するためには，統計量の漸近分布が未知パラメータに依存しないことと，高次のモーメントが存在することが必要である。

しかしながら，以上の結果を考えれば，上記のような条件が満たされない場合でも，平滑化ブートストラップ法を用いることにより，通常の漸近分布やブートストラップ法を用いるよりも正確な信頼区間を求めることができる可能性がある。

しかしながら，以上の結論は限定されたシミュレーション結果から得られたものであり，理論的な分析は未だ行われていない。したがって，全ての場合にこの方法が有効であるといえるわけではない。また，$\hat{h}_2$ というバンド幅は，未知の分布が正規分布であると仮定して得られたバンド幅にプラグ・インを用いて得られた値であり，それぞれの分布に適した値は個別に求めるべきかもしれない[1)]。これらの問題に関する分析は，今後の研究課題である。

注

1）カーネル密度推定量を用いた分布関数の推定量の MISE を最小にする値を，クロス・バリデーションにより求める方法を Bowman, Hall and Prvan（1998）が提案している。また，密度関数の推定量の MISE を最小化する値をクロス・バリデーションによって求める方法が Hall（1983），Stone（1984），Härdle, Hall and Marron（1988）等によって提案されている。

参 考 文 献

Beran, R., 1988, “Prepivoting Test Statistics: A Bootstrap View of Asymptotic Re nements,” *Journal of the American Statistical Association*, 83, 687-697.

Bowman, A., P. Hall, and T. Prvan, 1998, “Bandwidth Selection for the Smoothing of Distribution Functions,” *Biometrika*, 85, 799-808.

De Angelis, D., and G. A. Young, 1992, “Smoothing the Bootstrap,” *International Statistical Review*, 60, 45-56.

Efron, B., 1979, “Bootstrap Methods: Another Look at the Jackknife,” *Annals of Statistics*, 7, 1-26.

El-Nouty, C., and A. Guillou, 2000, “On the Smoothed Bootstrap,” *Journal of Statistical Planning and Inference*, 83, 203-220.

Hall, P., 1983, “Large Sample Optimality of Least Squares Cross-Validation in Density Estimation,” *The Annals of Statistics*, 11, 1156-1174.

Hall, P., 1992, *The Bootstrap and Edgeworth Expansion*, Springer-Verlag Inc.

Hall, P., T. J. DiCiccio, and J. P. Romano, 1989, “On Smoothing and the Bootstrap,” *The Annals of Statistics*, 17, 692-704.

Härdle, W., P. Hall, and J. S. Marron, 1988, “How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum?,” *Journal of the American Statistical Association*, 83, 86-95.

Li, Q., and J. S. Racine, 2007, *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Polansky, A. M., and W. R. Schucany, 1997, “Kernel Smoothing to Improve Bootstrap Confidence Intervals,” *Journal of the Royal Statistical Society, Series B, Methodological*, 59, 821-838.

Rosenblatt, M., 1956, “Remarks on Some Nonparametric Estimates of a Density Function,” *The Annals*

of Mathematical Statistics, 27, 832-837.

Silverman, B. W., 1986, *Density Estimation for Statistics and Data Analysis, Chapman* & Hall Ltd.

Silverman, B. W., and G. A. Young, 1987, "The Bootstrap: To Smooth or Not to Smooth?," *Biometrika*, 74, 469-479.

Stone, C. J., 1984, "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *The Annals of Statistics*, 12, 1285-1297.