化石　97，23–37，2015

Fossils
The Palaeontological Society of Japan

解　説

ノンパラメトリックな多峰性検定―Silvermanの検定―とその古生物学への導入

楠橋　直・岡本　隆
愛媛大学大学院理工学研究科

A nonparametric multimodality test—Silverman's test—and its introduction into paleontology

Nao Kusuhashi and Takashi Okamoto
Graduate School of Science and Engineering, Ehime University, Ehime 790-8577, Japan (nkusu@sci.ehime-u.ac.jp; okamoto@sci.ehime-u.ac.jp)

はじめに

統計学は得られたデータを如何に客観的に判断するかという指標を与えるものであり，統計学的手法は自然科学をはじめ様々な分野の研究で使われている．統計学的仮説検定（以下では検定と呼ぶ）はそのような手法のうちの1つであり，日本の古生物学においても，特に速水（1969）や速水・松隈（1971）がいくつかの基本的な検定法を導入して以来，広く研究に使われてきた．近年のコンピュータの発達に伴ってその手法はいよいよ多様化し，今後はいっそう高速かつ容易に様々な検定結果を手にすることができるようになるだろう．一方で危惧が無くもない．客観性・汎用性を目指してデータを画一化すれば，必然的に個性が失われる．その過程で付随する情報は省かれ，自信のあるデータもそうでないものもみな同じ顔つきになってしまうだろう．そして入力を終えボタンか何かを押せば何がしかの結果が自動的に返って来る．中でどのような計算がされているのかなど理解する余裕もなく，ただブラックボックスから出てきたものを，また無批判に人に伝える．それで良いのだろうか．

近年検定の安易な使用には警鐘が鳴らされている（例えばJohnson, 1999）．検定に対する考え方は様々だとしても，理論を理解せずに検定をおこなうことによる誤用は明らかに問題で，その結果が示すのはデータや母集団の性質ではなく，論文著者の「統計力」のなさだけである．統計学の進歩とともに，我々はもっと「統計力」を身につけるべきであろう．

本稿で紹介するのは，最近著者の一人楠橋の出会った1つの検定法である．その検定は，楠橋にとっては「斬新な」アイディアに立脚しながらも，手法はきわめてオーソドックスで理解し易くまた有用そうでもあった．そこで同僚の岡本を誘ってプログラムを作り，いくつもの例について解析をおこない，理論や実用性，そして欠点に関して議論を重ねた．結果，この検定法は古生物学の分野に広く紹介するに値するという結論に至った．それがSilvermanの検定（Silverman, 1981, 1983）である．

上述のように，統計学的手法はその理論を理解しないまま安易に利用すべきではない．そこで本稿では，読者が理解し易くするため，まず一般的な検定の考え方を述べ，それと対応付けてSilvermanの検定を解説する．その後古生物学への適用例を示し，最後にこの検定の問題点を挙げる．また著者の一人岡本が作成した検定用プログラムを公開（「化石」電子版のsupplement materials）するが，そのプログラムも検定を理解し易くするために工夫してある．

多峰性分布の検定

現在日本の古生物学で一般に用いられている検定は，主に単峰性分布を仮定したものであるが，実際には，古生物学においてはしばしば計量の分布が多峰性を示す場合がある．例えば何かの標本群について個々の標本のある長さを計測したとき，計測値の分布が複数のモード（以下本稿ではモードとは確率密度関数の極大の意味で用いる）をもつ場合があるかもしれない．複数のモードをもつ連続分布（多峰性分布）は，複数の単峰性分布が足し合わされたものであると解釈することができる．したがって計測値の分布が複数のモードをもつ場合，母集団には複数の部分母集団が含まれている可能性を指摘できる．母集団に複数の部分母集団が含まれているとすると，それは例えば成長段階の違いによるものかもしれないし，性的二型によるものかもしれない．場合によっては種の違いを示しているのかもしれない．そのため計測値の分布が複数のモードをもつことは，古生物学者にとっては好都合な「おいしい」ことで，複数のモードをもつことをできるだけ客観的に（見える方法で）示すことができ

れば，さらに次の議論へと繋げ易くなる．

多くのデータがあり，その分布が明瞭な複数のモードをもつ場合，ヒストグラムを作成するだけでも，そのことを示すには十分であろう．しかし化石の研究ではそう都合良く多くのデータが採れる場合ばかりではないし，また小さくて評価の難しいモードが出て来る場合も考えられる．そうなると，階級の幅や開始位置によって形の変わるヒストグラムでは，複数のモードをもつことを主張しにくい．そのようなとき，もし統計学的に計測値の分布が複数のモードをもつことを言えたならば，単にヒストグラムだけで示すよりも説得力が増すだろう．

複数の単峰性分布が足し合わさった分布については，個々の分布の母数や足し合わさった分布の個数を推定するために，古典的なところではPearson（1894）など以来，正規混合分布を中心に多くのパラメトリック，セミパラメトリックな手法での研究がおこなわれ，様々な分野に応用されてきた（例えばRender and Walker, 1984; Everitt, 1996; McLachlan and Peel, 2000; 金田・新居, 2009a, bなどを参照）．また同時に，ある分布が複数の正規分布の足し合わされたものなのか否か，さらには足し合わされた正規分布の個数を調べる検定についても研究が進められてきている（例えばAitkin and Rubin, 1985; McLachlan, 1987; Lo *et al.*, 2001）．地球科学の分野でも，例えばフィッション・トラック年代測定では，複数の正規分布が足し合わされている場合に，個々の正規分布に分解する方法が考案されている（例えばBrandon, 1992）．

しかしながら古生物学で扱うものでは，分布の型を仮定できない場合が少なくない．したがってノンパラメトリックな手法のほうが汎用性は高い．ノンパラメトリックな手法では，足し合わさった単峰性分布の型を仮定しないから，足し合わさっている分布の個数を厳密に知ることは，パラメトリックな手法と比べてより難しい．しかし，多峰性分布であるか否かを検定することは可能で，いくつもの検定が考案されてきた．一変量での検定の代表的なものとしては，本稿で紹介するSilvermanの検定のほかにも，Good-Gaskinsの検定（Good and Gaskins, 1980），dip検定（Hartigan, 1985; Hartigan and Hartigan, 1985），Wongの検定（Wong, 1985）などがあり，またより視覚的にモード数を知るためにはmode tree（Minnotte and Scott, 1993）とそれを発展させたmode forest（Minnotte *et al.*, 1998），そしてSiZer（Chaudhuri and Marron, 1999）を利用する方法がある．多変量での検定も多く提案されている（例えばMüller and Sawitzki, 1991a, b; Hartigan and Mohanty, 1992; Rozál and Hartigan, 1994; Polonik, 1995; Burman and Polonik, 2009）．

本稿ではこれらの検定のうちSilvermanの検定を紹介する．それはこの検定の理論が平易で理解し易く，かつ十分実用に堪えるものだと考えたからである．

検定とは

Silvermanの検定はノンパラメトリックな検定であるが，その手順は一般的によく知られているパラメトリックな検定のそれと対応させて理解することができる．そこでSilvermanの検定法を紹介するに先立って，ごく簡単な事例に即して一般的な検定ではどのような物の考え方をするのかを示しておきたい．一般的な検定を十分に理解している読者は，ここを読み飛ばしてもらって差し支えない．

帰無仮説と対立仮説

あまり穏当な例えではないが，検定という行為を一言で例えるなら，それは刑事裁判に似ている．この時被告人は，本来,「無実」か「犯人」かの2通りしか有り得ないのだが，裁判の本質を知るには,（1）無実,（2）証拠不十分な犯人,（3）明らかな犯人，の3通りを考えた方がわかり易い．検察は，様々な証拠をあげて犯行の立証を試みるのであるが，これが成功すれば被告人は（3）と認定され,「有罪」と判決される．一方，成功しなかった場合には被告人は「無罪」となるのだが，これは必ずしも（1）の無実を意味しない．その中には，実際犯行をおこなったけれど証拠不十分でそれが立証されなかった（2）の場合も含まれているのである．

検定の論理構造もこれと相似的で，検定試料は事実上,（1）差がない,（2）差があるけれど判らない,（3）差がある，の3通りにカテゴライズされる.（1）を（2）および（3）から峻別できるならそれに越したことはないのだが，全知全能でもない限りそれはほとんど不可能なことである．ないということの難しさは，古生物に関わる者であれば容易に理解できるであろう．そこで,「もし差がないとしたら」という前提に立って，それでは説明できないことを立証することで，せめて「明らかな差がある」事象を他から区別しようというのが検定である．このような目的でおいた前提を帰無仮説（H_0）という．そして帰無仮説の棄却に成功すれば対立仮説（H_1）が支持され，試料は「有意の差あり」と積極的に認定されることになる．しかしもし帰無仮説を棄却できなかった場合，注意しなければならないのは，帰無仮説が支持された訳ではない点である．この場合試料は,（1）だけでなく（2）である可能性も排除できない．そのため,「有意の差があるとはいえない」とか「有意の差が認められない」などという弱い表現にならざるを得ない．間違っても「差がない」などと断言してはいけない．

検定では，帰無仮説が成り立つ確率が予め設定した十分小さな確率（有意水準）よりも小さいとき，その帰無仮説を棄却する．つまりある事象が帰無仮説の下では非常に小さい確率でしか起こり得ないから，帰無仮説は成り立たないと結論するわけである．しかしこのことは同

時に，その事象は非常に小さい確率ではあるが，帰無仮説の下でも起こり得ることを示している．したがって，検定においては，実は帰無仮説が成り立っているにもかかわらず，帰無仮説を棄却してしまう誤りが存在する．これを第一種過誤と呼ぶ．上の裁判の例で言えば，無実であるにもかかわらず有罪と判決してしまう冤罪に相当する．また検定には逆に，実は帰無仮説が成り立っていないにもかかわらず棄却しない誤りもあり，これは第二種過誤と呼ばれる．第一種過誤の起こる確率と第二種過誤の起こる確率は逆相関関係にある．また第一種過誤の確率は有意水準に等しく，有意水準は自ら設定するものだからその確率を小さくすることはできる．一方で第二種過誤の確率は検定法の性質やデータ数など様々な要因によって変化する．通常は有意水準を十分小さく設定するから，「有意の差がある」ことは積極的に認定できるが，第二種過誤の確率は大きくなり得るので，帰無仮説を棄却できない場合は控えめな表現となるのである．

仮想的な問題設定

ごく簡略化した形で検定の手順を説明するために，いま，M君が1個体だけ化石を採集し，それが種Aとは異なる種なのではないかと考えた―そんな架空の場面を想定しよう．ただし，種Aのある形質に関する計測値xは連続変量であり，平均値$\mu=10$，標準偏差$\sigma=2$の正規分布に従うことが既に知られているものとする．そこでM君は化石を計測して同じ形質について$x=14$という数値を得た．果たしてこの標本は，統計学的にみて種Aとは異なるものと見做して良いのだろうか．

検定をするにあたってまずすべきことは，帰無仮説を立てることである．この場合帰無仮説は，「M君の化石は種Aである」ということになるだろう．もう少し正確に表現するなら，「この標本は種Aという無限母集団から無作為に由来したものである」となる．上にも述べたようにこれを立証することは不可能である．なぜなら，いくら標本の数値が種Aに近くても種Aとは異なる母集団から由来した可能性を排除できないからである．しかし，その数値が種Aとして期待されるものから離れていた時には，帰無仮説をある確からしさで否定することができる．そうした場合，同じ確からしさで対立仮説である「M君の化石は種Aではない」が支持されるのである．

母集団の確率密度分布

帰無仮説が決まったら，それを仮定したときの無限母集団の確率密度分布を想定する必要がある．視覚的なイメージとしては，ある計量を無限回計測した結果を限りなく細かい区画のヒストグラムで表したものが確率密度分布である．ただしその計量のとり得るあらゆる値に関する累積確率密度は1だから，ヒストグラムの面積は1になるように基準化する．このヒストグラムの上端を滑らかに繋いだ曲線は，その計量についてそれぞれの数値xがどれだけの確率密度$f(x)$で現われるかを示しており，これを確率密度関数と呼ぶ．つまりヒストグラムにおいて，各階級の幅をw，データ総数をnとしたとき，各階級に高さ$1/nw$のブロックを度数に応じて積み重ねたと考えれば，それは確率密度分布の概形になっているわけである．

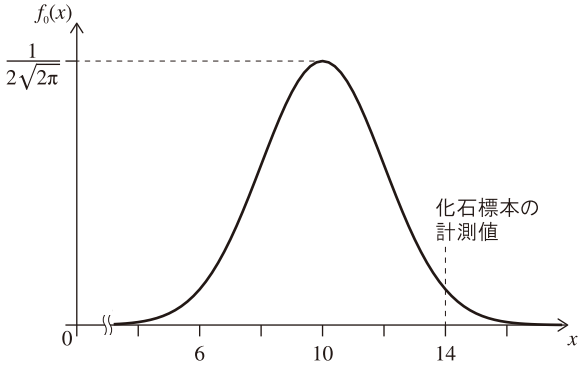


図1．帰無仮説を仮定したときに推定される母集団の確率密度分布．本文の例の場合，種Aのその形質に関する計測値は平均値10，標準偏差2の正規分布に従う．

以下では，帰無仮説を仮定したときの母集団の確率密度分布をf_0，その確率密度関数を$f_0(x)$と書くことにする．いまの場合，種Aにおける件の形質については，平均値と標準偏差が既知の正規分布に従うことがわかっているので，確率密度関数は次のように表される（図1）．

$$f_0(x)=\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-10)^2}{8}}$$

検定統計量

次に適当な検定統計量を決めて帰無仮説が正しい場合の検定統計量の確率密度関数を求める．検定統計量というのは，実際の標本（以下では経験標本と呼ぶ）とf_0から抽出された仮想標本とを何をもって比較するかという，その量のことである．この例の場合は，標本の計測値とf_0の平均値μとの差を求め，それを標準偏差σで除したもの$z=|x-\mu|/\sigma$が良いだろう．したがって，経験標本から計算される検定統計量は$z=2$である．今度はこれを変数と見做して確率密度関数$g_0(z)$を求める．ここでいう確率密度関数$g_0(z)$は，f_0から標本と同じ個数（1個）の抽出をおこなったときに計算される確率密度をzで表わしたもので，いまの場合は，平均0，標準偏差1の正規分布曲線の右半分を高さだけ2倍したものになっている（図2a）．

$$g_0(z)=\frac{2}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}\quad(z\geq 0)$$

この例では，統計量の確率密度関数$g_0(z)$が母集団分布の確率密度関数$f_0(x)$とたまたま同種類の関数になったが，両者は基本的に異なるものであることに注意されたい．

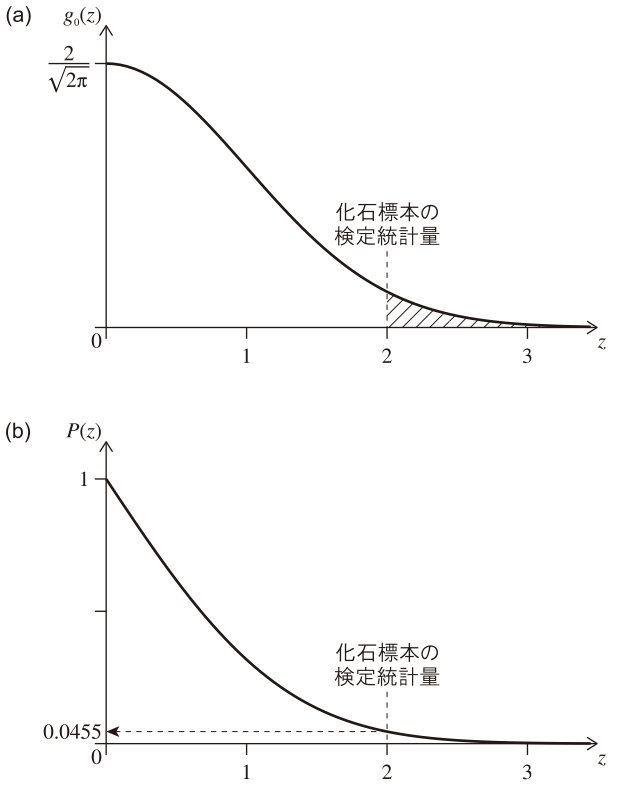


図2．帰無仮説を仮定した場合に推定された検定統計量の確率密度関数（a）とP値（b）．確率密度関数においては検定統計量が観測値以上に極端な値をとる確率P値は斜線部の面積になる．P値はzの単調減少関数になっている．

$g_0(z)$は，F検定ならF分布，t検定ならt分布，χ^2検定ならχ^2分布…というように，実際には様々なものがある．

*P*値を求める

統計量のP値というのは，帰無仮説を前提とした時に観測値がzよりも極端な値をとる確率であり，zの関数$P(z)$として表すことができる．$P(z)$は一種の累積関数で，値域が$0 \le P(z) \le 1$の単調減少関数となる．視覚的には確率密度関数$g_0(z)$と横軸とで囲まれた部分のうち定義域がzより大きい領域の面積を表わしていて，$g_0(z)$をzから∞まで定積分したものになる（図2b）．

$$P(z) = \int_z^{\infty} g_0(z)dz$$

実際にP値を計算で求めるのはなかなか厄介なので，通常は予め結果が計算されている数表を用いて見積もることが多い．いまの例では，観測値の統計量は$z=2$であるから，標準正規分布の数表で$z=2.00$の欄を読むと$p=0.02275$という数値が得られる．ただしこの数値は標準正規分布において，与えられた域値よりも大きい側の部分の面積である．右半分しかない$g_0(z)$は標準正規分布の2倍の高さをもっているから，$P=2p=0.0455$と見積もることができる．すなわち，M君の標本が種Aである可能性は，この形質で見る限り4.55％ということになる．

ところで，推測統計学では一般に5％より小さい確率を「ありそうもないこと」と見做し，帰無仮説が95％以上の確率で棄却できる時に「有意の差がある」と表現する．したがって，M君の標本は種Aとは形態的に有意の差があり，種Aではありそうもないと結論付けることができる（これがもし仮に$P \ge 0.05$だったら，有意の差があるとはいえず，種Aである可能性も否定できないなどと表現することになる）．このときの5％が上述した有意水準である．5％というのは全く人為的に決めた数値であるから，「こういう場合にこのような表現が使える」程度に思っていれば良い．より本質的なのはあくまで確率P値の数値である．また，有意水準には，求められている正確さによって（第一種過誤の起こる確率を小さくするために）更に厳格な1％や0.1％が用いられる場合もある．

ブートストラップ検定

さて，統計量のP値は，通常，数表を用いて求めると述べたが，もし数表が無かったらどうしたら良いだろう．これまで述べてきた手順を見ればわかるように，検定の際には$P(z)$は必ずしもその全体像を知らなければならない訳ではない．観測された特定の検定統計量に対応するP値が求まれば事足りるのである．そこで1つの解決策として実験的な手法が有り得る．すなわち，実際の標本と同じ個体数だけf_0から抽出し，そこから計算された検定統計量を実際に観測されたものと比べるのである．1度や2度の抽出では結論は出ないが，これを何回も繰り返したら近似的にP値を見積もることができるだろう．これにはブートストラップ法を援用するのがよい．

ブートストラップ法は，Efron（1979）によって導入された，母集団の性質を評価するための手法である．観測されたn個のデータから，無作為にn個復元抽出して疑似データ（ブートストラップ標本）を生成し，その疑似データから推定量を計算する，ということを何度も繰り返すのがブートストラップ法の特徴である．ここで復元抽出とは抽出の際に同じデータが何度重複して選ばれても構わないという意味である．

ブートストラップ法を用いた検定（ブートストラップ検定）では，一般に，データ$x_i (i=1, 2, \ldots, n)$から求めた検定統計量tと，復元抽出したブートストラップ標本$x_i^* (i=1, 2, \ldots, n)$から求めた$t^*$との間に特定の関係が成り立つかどうかを調べて$P$値を見積る．例えば，帰無仮説$H_0$が成り立たないならば，$t$は$H_0$が成り立つときよりも大きな値をとることが期待されるなら，

$$P_{\mathrm{boot}} = \Pr_{H_0}\{t^* \ge t\}$$

である．ただしP_{boot}はブートストラップ検定でのP値，Pr{A}は条件Aを満たす確率である．実際にはデータx_iからブートストラップ標本x_i^*を復元抽出してt^*を求める

ということをB回繰り返し，P_{boot}値を

$$\hat{P}_{\mathrm{boot}}=\frac{\#\{t^*\geq t\}}{B}$$

で近似する（Efron and Tibshirani, 1994）．ただし#{A}は条件Aを満たす回数である．ブートストラップ法や検定については様々な参考書が出版されているので（例えばEfron and Tibshirani, 1994; Davison and Hinkley, 1997; 汪・桜井, 2011），詳細はそれらを参考にされたい．

基本的なブートストラップ法では，上述のように，実際の標本の分布（経験分布）そのものを母集団の分布f_0として与えているため，当然それは離散的な分布構造をしている．復元抽出なので減ることはないが，このやり方ではブートストラップ標本としていくら抽出をおこなっても，すでに取った事のあるデータしか現れない．これでは本来滑らかな分布であるはずの母集団からの抽出標本として必ずしも妥当とはいえないだろう．そこで，ただブートストラップ標本を抽出する代わりに，毎回の抽出ごとに多少のランダムノイズを加える方法が考案されており，このようにして得られた標本を平滑化ブートストラップ標本（Efron, 1979）と呼ぶ．この操作は，実はf_0として経験分布よりも滑らかな分布を想定しようとしているに等しく，後に述べるカーネル密度推定（kernel density estimation）の援用に外ならない．

Silvermanの検定では帰無仮説が正しい場合の検定統計量について確率密度関数$g_0(z)$が一般化できないので，P値に関して予め計算しておくことができない．そのため平滑化ブートストラップ標本を用いてその都度統計量のP値を見積もるのである．コンピュータの発達した今日ならではの力技といえるだろう．

Silvermanの検定

Silvermanの検定の原理を知るにはカーネル密度推定を理解しなければならない．そこでまずカーネル密度推定について紹介した後，Silvermanの検定の論の進め方を，上で述べた一般的な検定と対応付けながら解説する．なお，この検定については，和文では樋田（1999）が簡単に説明している．また以下では基本的なSilvermanの検定についてのみに限って紹介するので，Silvermanの検定を発展させる試みについては，Chan and Tong（2004），Hall *et al.*（2004），Ahmed and Walther（2012）などを参考にされたい．

カーネル密度推定

カーネル密度推定はノンパラメトリックな確率密度推定の代表的手法の1つである．この方法は，Rosenblatt（1956），Parzen（1962）らによって一変量の場合について研究され，後にCacoullos（1966）やEpanechnikov（1969）らによって多変量へと拡張された．Silvermanの検定は一変量の場合の検定であるから，以下では一変量のカーネル密度推定について簡単に解説する．その詳細や多変量のカーネル密度推定についてはSilverman（1986）やBishop（1995）などをはじめ多くの文献で解説されているのでそちらを参考にされたい．

同一分布に従う互いに独立なデータ$x_i\,(i=1, 2, \ldots, n)$が得られたとき，それらの従う確率密度関数$f(x)$を推定したい．上で述べたように，ヒストグラムを作成すれば$f(x)$の概要を掴むことはできる（図3a）．そしてカーネル密度推定は，実はヒストグラムから$f(x)$を推定する方法と基本的には変わらない．しかしヒストグラムは階級を区分したことにより，$f(x)$とは無関係に不連続で，しかも階級の取り方によってその形が大きく変化するという問題がある（例えばFisher, 1989, figure 1）．そこで，データを階級に振り分けるのではなく，ある一定の幅をもつ領域を動かしながら，その領域の中に観測値が入る確率から$f(x)$を推定する，というのがノンパラメトリックな密度推定の基本的な考え方の1つであり，カーネル密度推定もまたそれに基づいて$f(x)$を推定する．

ごく単純化して言えば，一次のカーネル密度推定とは，個々のデータ点x_iを中心として，$\int_{-\infty}^{\infty}K(u)\,\mathrm{d}u=1$であるようなカーネル関数（kernel function）と呼ばれる関数$K(u)$を縦横に一定の割合で伸縮したものをn個置いていくことで，確率密度関数を推定する方法である．最も簡単なカーネル関数の例として，

$$K(u)=\begin{cases}1, & |u|<1/2\\ 0, & (\text{otherwise})\end{cases}$$

という関数を考える．このカーネル関数はParzen窓関数（Parzen window function）と呼ばれ，幅が1，高さが1（したがって面積が1）の矩形をしている．これを幅hに引き伸ばし，高さを$1/nh$倍したブロック状の要素を，各データ点を中心に置いていけば，また全体の面積が1のヒストグラムに似た外形となる（図3b）．ヒストグラムと違って階級は区分されていないから各階級の度数というものはないが，その代わり$\Sigma K\{(x-x_i)/h\}$はxにおいていくつのブロックが重なっているかを示している．そのとき，その形が確率密度分布の概形となるのはヒストグラムと同様である．したがって$f(x)$は，

$$\hat{f}_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K(u)\quad\text{ただし}\ u=\frac{x-x_i}{h}$$

と推定できる．この$\hat{f}_h(x)$をカーネル密度推定量（kernel density estimator）と呼ぶ．バンド幅（bandwidth）hはカーネル関数の横への広がりの程度を示している．

カーネル関数にParzen窓関数を使った上の推定だと，ヒストグラムと違って階級区分は必要なくなる．しかし，xとx_iとの距離に関わらず一律に重みづけがされており，求められた$f(x)$は依然不連続なままであるという問題が残る．ならばParzen窓関数の代わりに滑らかなカーネル関数を用いれば良い．一変量の場合，カーネル関数には

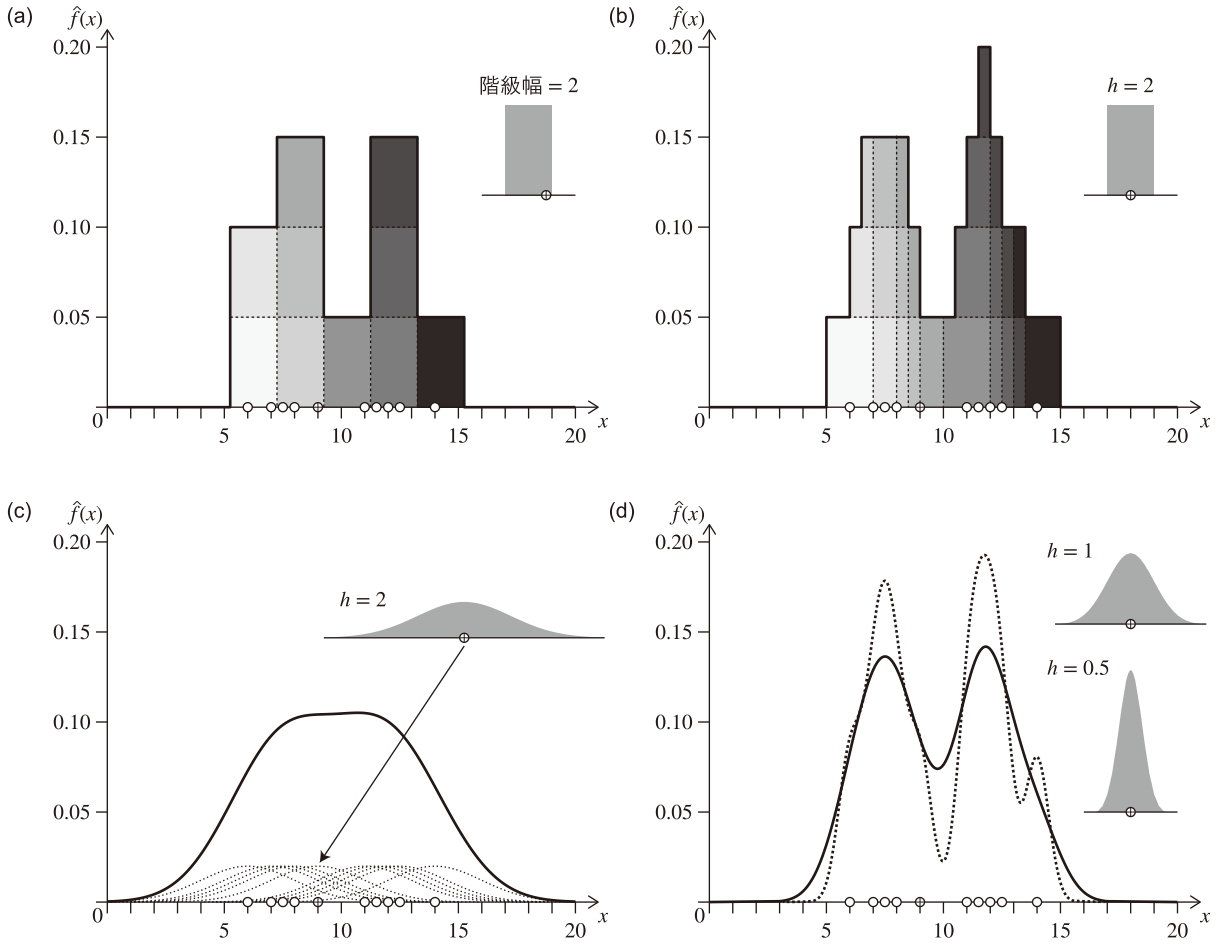


図3．10点の架空データ（x軸上の白丸）に関するヒストグラム（a）とカーネル密度推定（b–d）．太線（実線および点線）は推定される確率密度関数を表わし，右上には確率密度関数を構成する要素と，データ点との関係を示している．（a）階級幅2のヒストグラム．要素の階級区切りは予め決まっている．（b）バンド幅2のParzen窓関数をカーネル関数に用いたカーネル密度推定量．概形はヒストグラムに似るが，階級区切りは必要ない．（c）バンド幅2のGaussianカーネルを用いたカーネル密度推定量．Gaussianカーネルではバンド幅が標準偏差となるため，確率密度関数はより横に広がった形となる．（d）バンド幅1（太線）と0.5（点線）のGaussianカーネルを用いたときのカーネル密度推定量．バンド幅が小さくなると，確率密度関数の極大数が増える．

例えばGaussianカーネル（標準正規分布関数）

$$K(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$$

や，Epanechnikovカーネル

$$K(u)=\begin{cases}\dfrac{3}{4}\left(1-\dfrac{1}{5}\mathrm{u}^2\right)/\sqrt{5}, & |\mathrm{u}|<\sqrt{5}\\ 0, & \text{(otherwise)}\end{cases}$$

などが用いられる．

Silvermanの検定では，母集団の確率密度関数を推定する際に，Gaussianカーネルを使ってカーネル密度推定をおこなう．この場合，カーネル密度推定量は，標準正規分布の横縦を，それぞれ，h倍および$1/nh$倍にした要素を個々のデータ点x_iに置き，それらすべてを足し合わせたものとなる（図3c）．

カーネル密度推定量は一見してわかるように，バンド幅hに大きく依存する．すなわちhはカーネル密度推定量の平滑化パラメータである．hが大きくなると全体がスムーズになり，hが小さくなると個々のデータの影響を受け易くなる（図3d）．つまりhはヒストグラムにおける階級幅と同じような働きをする．

カーネル関数にGaussianカーネルを使った密度推定では，$\hat{f}_h(x)$のもつ極大の数$N(\hat{f})$はhの非増加関数になることがSilverman（1981）によって証明されている．つまり，バンド幅（ここでは要素の標準偏差）を大きくしていった時に$N(\hat{f})$が増えることは決してない．これはSilvermanの検定の根幹を支える重要な性質である．そしてSilvermanの検定の巧妙な点は，この性質を利用して最も妥当な$N(\hat{f})$はいくつと考えればよいかを確率的に判断しようとしたところにある．なお，カーネル密度推定にGaussianカーネル以外を用いると，$N(\hat{f})$がhの非増加関数になることは保証されない（Minnotte and Scott, 1993; Hall *et al.*, 2004）．

帰無仮説と対立仮説

Silvermanの検定では，同一分布に従い互いに独立なデータが得られたとき，そのデータの母集団がk個より多くのモードをもつか否かを検定する．そこで帰無仮説H_0^kは「母集団はたかだかk個のモードしかもたない」とする．したがって対立仮説H_1^kは「母集団はk個よりも多くのモードをもつ」である．

母集団の確率密度分布

次に帰無仮説H_0^kが正しい場合の母集団の確率密度関数$f_0^k(x)$を推定する．この関数は，当然，k個の極大を持っているものでなければならないが，同時に，そこから得られた経験標本中に$k+1$個目のピークが最も現れ易いものでもある必要がある．

実際，Silvermanの検定では，$f_0^k(x)$を，データに対してGaussianカーネルを使ったカーネル密度推定を用いることで推定する．このとき，前述のように，この関数における極大の数$N(\hat{f})$はバンド幅hの増加にともなって決して増加しないことから，$N(\hat{f})=k$であるような最小のバンド幅$h_{k.crit}$（k臨界バンド幅，k-critical band width）を定義できる．したがってこの$h_{k.crit}$を用いて推定されるカーネル密度推定量

$$\hat{f}_{h_{k.crit}}(x)=\frac{1}{nh_{k.crit}}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h_{k.crit}}\right)$$

は，極大数がkとなるギリギリの$\hat{f}_h(x)$であり（図4），hが$h_{k.crit}$より少しでも小さくなると，$\hat{f}_h(x)$の極大数は$k+1$となる．逆にhをこれより大きくとれば，データをより平滑化して確率密度関数を推定することになるから，そこから得られた経験標本中に余分なピークが現れることはより期待しにくくなるだろう．したがって，帰無仮説に従う確率密度関数としては，$\hat{f}_{h_{k.crit}}(x)$が最もデータに則したものであり，これを$f_0^k(x)$の原型とするのである．

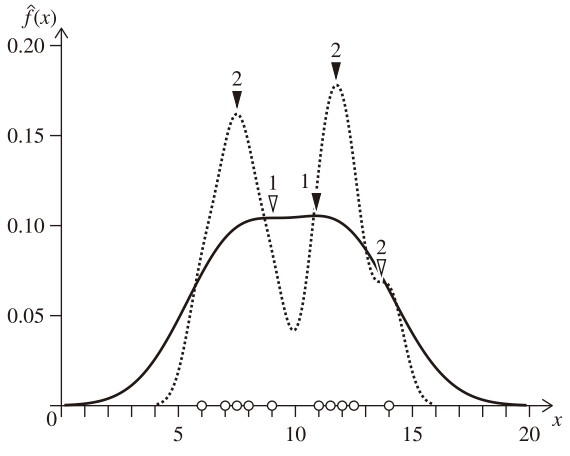


図4．図3のデータに関して$k=1$および2のときの臨界バンド幅でおこなわれたカーネル密度推定．それぞれ黒矢印は極大の位置を，白矢印はギリギリ現れない潜在的な極大（$\hat{f}'(x)=0$だが$\hat{f}''(x)$の符号は変わらない点）の位置を示す．

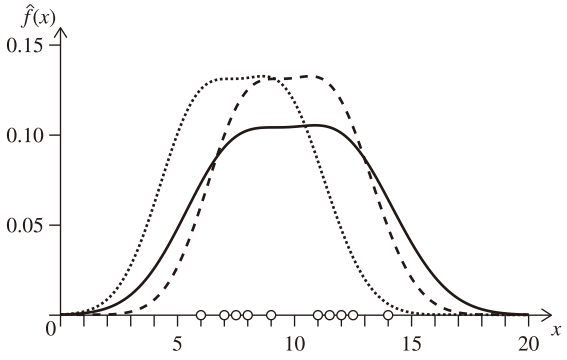


図5．Silverman（1981）と，Efron and Tibshirani（1994）のそれぞれの方法でおこなった分散調整の結果．どちらの方法でも$\hat{f}_{h_{k.crit}}(x)$の分散が経験標本の分散に一致するように補正される．Silverman（1981）の方法で補正したもの（点線）は補正前（実線）に比べて平均がずれるが，Efron and Tibshirani（1994）の方法（破線）だと平均はずれない．

しかしながら，$\hat{f}_{h_{k.crit}}(x)$をそのままの形で$f_0^k(x)$として採用するには問題が1つある．$\hat{f}_{h_{k.crit}}(x)$は，各要素についてバンド幅分の分散を人為的に付与しているので，全体の分散が経験標本の分散s^2よりも$1+h_{k.crit}^2/s^2$倍だけ大きくなってしまっているのである（例えばJones, 1991）．Silvermanの検定では，後で述べるように，$f_0^k(x)$からブートストラップ抽出によってデータを取り出して経験分布と比べる．このとき検定統計量として分布の形を直接比べる指標があるなら，あるいはこのままでも問題ないのだが，実際に使用している指標は標本のばらつきに強く依存してしまう．そこで$f_0^k(x)$は，$\hat{f}_{h_{k.crit}}(x)$の分散を経験標本の分散に一致させるよう補正したものとする（Silverman, 1981; Efron and Tibshirani, 1994）（図5）．

検定統計量

Silvermanの検定における検定統計量はk臨界バンド幅$h_{k.crit}$である．カーネル密度推定量$\hat{f}_h(x)$の極大数$N(\hat{f})$がバンド幅hに依存するわけだから，検定統計量をバンド幅とするのはわかり易い．仮に$h_{k.crit}$が“大き過ぎ”れば，それは本来$N(f)>k$である母集団分布を無理やり滑らかにして$k+1$個目の極大を消し，$N(\hat{f})=k$と見做していることを意味する．そのような場合には，$f_0^k(x)$から再抽出されたデータを同じバンド幅$h_{k.crit}$でカーネル密度推定しても$k+1$個以上の極大が再現されることは滅多に起こらなくなるだろう．つまり再抽出されたデータから計算される臨界バンド幅$h_{k.crit}^*$は$h_{k.crit}$よりも小さくなりがちになる．そこでブートストラップ検定により，$h_{k.crit}$と$h_{k.crit}^*$との大きさを比較し，$h_{k.crit}$が有意に“大きい”と言えたときに帰無仮説H_0^kを棄却するのである．

検定をおこなうにあたってはまず$h_{k.crit}$を求めなければならない．$h_{k.crit}$は例えば二分探索によって容易に求められる（Silverman, 1981）．なぜなら$N(\hat{f})$がhの非増加関数であることから，$N(\hat{f})>k$となるのは$h<h_{k.crit}$の場合

に限られるからである．

*P*値

Silvermanの検定では，平滑化ブートストラップ標本を用いたブートストラップ検定により，k臨界バンド幅$h_{k.crit}$が“大きい”かどうかを判断する．上述のように帰無仮説H_0^kが正しくなければ，$h_{k.crit}$は“大きい”はずであり，そのとき$h_{k.crit} \geq h^*_{k.crit}$となるであろう．そこでこの検定では

$$P^k_{\mathrm{boot}} = \Pr_{H_0^k}\{h^*_{k.crit} > h_{k.crit}\}$$

とする．

具体的なブートストラップ検定の手順としては，まず，n個のデータx_iからn個のブートストラップ標本x_i^*を復元抽出する．そしてそこから平滑化ブートストラップ標本

$$y_i^* = \frac{1}{\sqrt{1+h^2_{k.crit}/s^2}}(x_i^* + h_{k.crit}\varepsilon_i)$$

を得る（Silverman, 1981）．ただしε_iは標準正規乱数である．この平滑化ブートストラップ標本抽出の際に$(1+h^2_{k.crit}/s^2)^{-1/2}$によって$f_0^k(x)$の分散を経験分布の分散$s^2$と同じに調整しているわけである（図5）．Efron and Tibshirani（1994）は平滑化ブートストラップ標本を

$$y_i^* = \overline{x^*} + \frac{1}{\sqrt{1+h^2_{k.crit}/s^2}}(x_i^* - \overline{x^*} + h_{k.crit}\varepsilon_i)$$

として得るとしており，こちらを用いると，$f_0^k(x)$の平均値を変えることなく分散の調整ができる（図5）．ただしSilvermanの検定では極大の数とバンド幅だけが問題だから，そういう意味ではどちらでも同じことである．

こうして得られた平滑化ブートストラップ標本より，$N(\hat{f}^*_{h_{k.crit}})$を求める．なお現実の計算に際しては，上でも述べたように，$N(\hat{f}) > k$となるのは$h < h_{k.crit}$の場合のみだから，$N(\hat{f}^*_{h^*_{k.crit}}) > k$であれば$h^*_{k.crit} > h_{k.crit}$であることはわかるので，$h^*_{k.crit}$を求める必要はない．これを$B$回繰り返して，

$$\hat{P}^k_{\mathrm{boot}} = \frac{\#\{h^*_{k.crit} > h_{k.crit}\}}{B}$$

を求め，有意水準αで検定するなら，$\hat{P}^k_{\mathrm{boot}} < \alpha$で$H_0^k$は棄却される．なおSilverman（1981）ではSilvermanの検定をGood and Gaskins（1980）のコンドライト隕石のデータに適用しているが（Silverman, 1981, table 1），ここには$1-\hat{P}^k_{\mathrm{boot}}$の値が誤って$P$値として示されている（Silverman, 1986）．

モード数の推定

検定は有意水準αとして検定モード数$k=1$でまずおこない，帰無仮説H_0^1が棄却されたなら，$k=2$で検定し，というようにkを1ずつ増やしながら検定を繰り返す．しばしば複数の異なるkでH_0^kを棄却できない，あるいは一度棄却できなくなってもさらにkを増やすと再び棄却できるという状況が生じるが，$k=1$から順に検定し最初にH_0^kが棄却できなかったときのkを$\hat{f}(x)$の極大数とすれば良いとされる（例えばSilverman, 1981; Efron and Tibshirani, 1994）．ただしIzenman and Sommer（1988）はP値が一定の値（彼らの例では0.40）を超えるまで続けることを提案している．

Silvermanの検定の適用例

Silvermanの検定を古生物学の研究に適用してみる．ここでは手取層群桑島層産トリティロドン類の下顎頬歯頬舌径（$n=49$）と，Okamoto and Shibata（1997）で計測された蝦夷層群流矢層産ポリプチコセラス（*Polyptychoceras pseudogaultinum*）の最終隔壁の位置データのうち第2シャフト後半以降のもの（$n=133$）の2つ

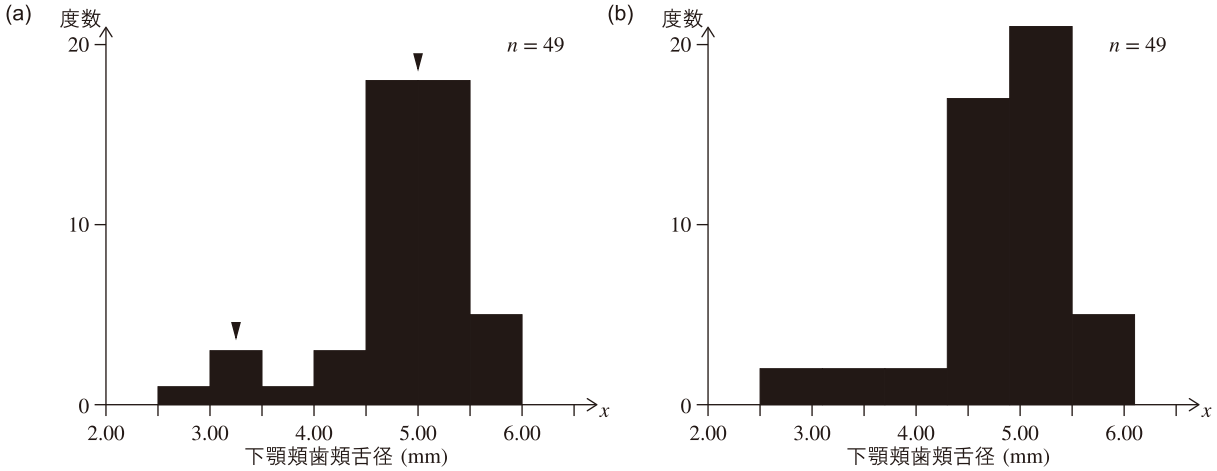


図6．桑島層産トリティロドン類の下顎頬歯の頬舌径に関するヒストグラム．(a) 最小の階級を2.50 mmからとし，階級幅0.50 mmで作成すると，2つのピークが見られる（矢印）．(b) しかし，階級幅を0.60 mmとすると，小さなピークははっきりしなくなる．そのためヒストグラムだけでは大きさの異なる2つのタイプがあることは強く主張しにくい．

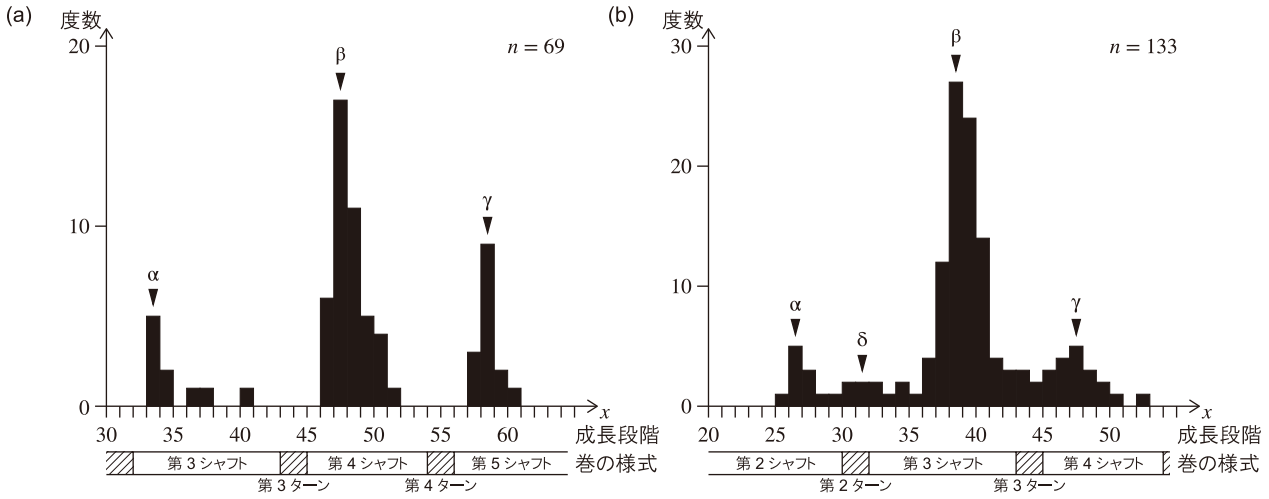


図7．流矢層産ポリプチコセラスについて計測された殻口まで保存された標本の殻口位置（a）と，最終隔壁の位置（b）のヒストグラム．Okamoto and Shibata（1997）で示されたもののうち第2シャフト後半以降のものに限って示した．α, β, γ, δはピークの位置．横軸には対応する殻の高さで規格化していった成長段階を取っている．

を例にとる．

桑島層産トリティロドン類の頬歯には，大きさが異なるもののそれ以外の形態ではほとんど区別のできない2つのタイプがあることが知られている（松岡, 2000）．楠橋の計測によれば，例えば下顎頬歯の頬舌径の場合，5 mm付近に見られる明瞭なピークのほかに，3.0～3.5 mm付近にも小さなピークをもつように見える（図6）．この小さなピークが統計学的に意味をもつならば，下顎頬歯には少なくとも2つの部分母集団の存在が示唆される．しかし，これらの頬歯は共産し，しかも頬歯サイズの分布型は未知だから，古生物を扱うに当たって普遍的に使われてきたこれまでの統計学的手法では，これらの計測値が2つの部分母集団に由来するか否かを評価するのは困難である．

ポリプチコセラスに関しては，Okamoto and Shibata（1997）によれば，計測された試料のうち，殻口まで保存された標本の殻口位置には，第3，4，5シャフトのやや前よりに互いに離れた明瞭なピークが見られる（図7a）．第2シャフト後半以降の最終隔壁の位置にもそれらに対応するであろう3つのピーク（α, β, γ）が見られるが，こちらは殻口位置のものほど明瞭ではない（図7b）．最終隔壁の位置は殻口位置と関係するはずだから，最終隔壁の位置も三峰であると考えられるが，Silvermanの検定では3つのピークを検出できるだろうか．

検定用プログラムの作成

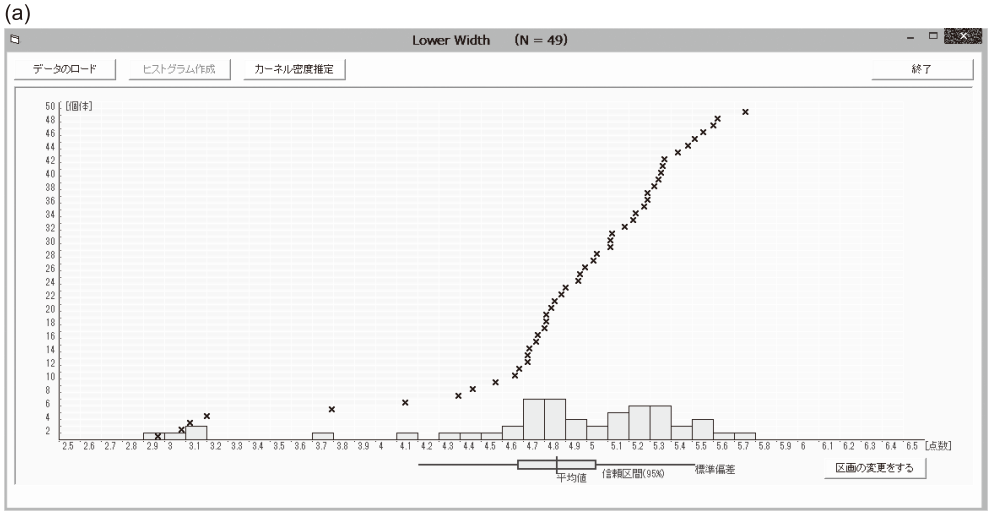
これらのデータに関するSilvermanの検定には，著者の一人岡本の作成したプログラムVISUAL-SILVERMANを使用した．例えば統計ソフトR（Ihaka and Gentleman, 1996; R Core Team, 2013）には‘silvermantest’パッケージ（Schwaiger and Holzmann, 2013）が公開されているなど，無料のアプリケーションも使えるのだが，それにもかかわらず敢えてプログラムを自作したのは，Silvermanの検定の中で実際にどのような操作をおこなっているのか，指摘されている問題点（後述）はどのような時に顕在化するのかなど，この検定についてより深い理解を目指したためである．作成したプログラムのソースコード（Visual Basic 6.0）と実行ファイルは，「化石」電子版においてsupplement materialsとして公開する．なお，統計ソフトRとの間で検定結果に齟齬がないことを確かめてある．k臨界バンド幅$h_{k.crit}$の計算では本プログラムの方がRよりもやや正確である．計算速度は劣るがインターフェイスなど利便性に関しては優れている．

このプログラムでは，(1) ヒストグラム作成，(2) 検定モード数をk (k = 1～4) としたときの$h_{k.crit}$の算出，(3) そしてブートストラップ検定と3段階で検定をおこなうようになっている．段階的におこなうことで，Silvermanの検定がより理解し易くなると考えたからである．

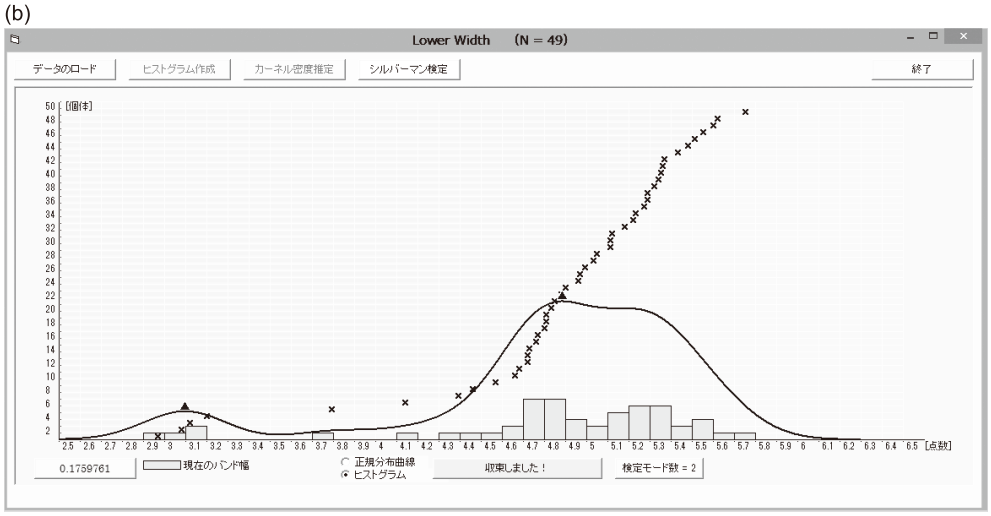
ヒストグラム作成の場面では，データを読み込み，適当に区画を調整した後にヒストグラムを作成し表示する（図8a）．もとのデータの型はテキストファイル形式で，これは各種のテキストエディタやMicrosoft Excelなどで作成することができる．ヒストグラムと同時に平均・標準偏差などパラメトリックな情報が表示され，次のカーネル密度推定に進むことができるようになる．この場面は，むしろ，データに入力ミスなどがないかを確かめるのが主な目的でSilvermanの検定自体とはほとんど関係がない．

カーネル密度推定の場面では，入力された検定モード数に対応する$h_{k.crit}$を計算する（図8b）．このとき，仮定したバンド幅でのカーネル密度推定量のグラフと検出された極大の位置が画面に表示される．初期設定ではバン

(a)



(b)



(c)

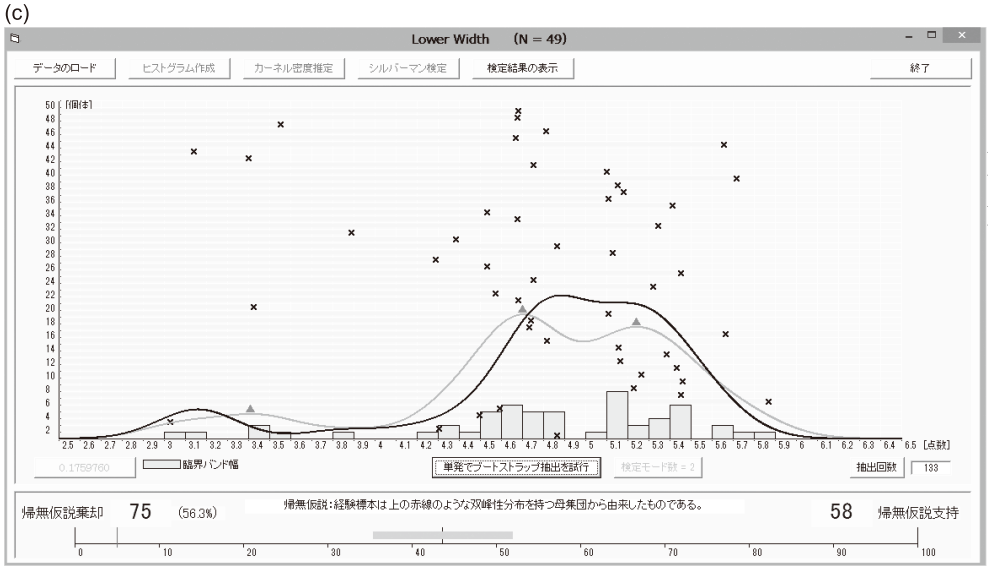


図8．自作プログラムVISUAL-SILVERMANによって桑島層産トリティロドン類の下顎類歯類舌径の分布を解析した例．(a) 値の順に並べられた計測値からヒストグラムを描き，パラメータを計算したところ．(b) $k=2$のときの臨界バンド幅の探索を終え，帰無仮説が主張する確率密度分布を求めたところ．2つのピークの位置は確率密度関数の曲線上に三角印で示される．(c) Silvermanの検定においてブートストラップ抽出をおこなっているところ．

表1．自作プログラムVISUAL-SILVERMANによるトリティロドン類の下顎類歯類舌径とポリプチコセラスの最終隔壁分布の検定結果．

トリティロドン類				
検定モード数（k）	1	2		
帰無仮説	分布は単峰性	分布は二峰性		
臨界バンド幅	0.461853	0.175976		
P 値	0.0165	0.4505		
判定	棄却	棄却不可		
ポリプチコセラス				
検定モード数（k）	1	2	3	4
帰無仮説	分布は単峰性	分布は二峰性	分布は三峰性	分布は四峰性
臨界バンド幅	2.690038	1.952924	1.599452	0.872359
P 値	0.0102	0.0290	0.0110	0.6126
判定	棄却	棄却	棄却	棄却不可

ド幅は短め（標準偏差の1/10）に設定されていて，そこから$h_{k.crit}$を探索していく．バンド幅が収束し$h_{k.crit}$が求まったら，次のSilvermanの検定に進むことができるようになる．

検定の画面では，はじめに帰無仮説に基づく$f_0^k(x)$が表示される．ボタンをクリックする度にブートストラップ検定がおこなわれ，帰無仮説に対する支持／棄却のスコア（P値）とその95％信頼区間が表示される（図8c）．また，抽出回数を入力することで，回数分の繰り返し自動計算も可能である．P値のエラーバーが$P=0.05$（画面表示では95％）のラインから外れるか，規定の回数分（500回）のブートストラップ検定がおこなわれたら，検定結果を表示できるようになる．

検定結果

1．トリティロドン類の類歯

下顎類歯の類舌径データから算出された臨界バンド幅$h_{1.crit}$，$h_{2.crit}$はそれぞれ0.461853，0.175976であった（表1）．$h_{1.crit}$のときは5 mm付近の大きなピークが，$h_{2.crit}$のときはさらに3.0～3.5 mm付近のピークが極大として現れていて（図8b），いずれもモードとして期待されていたものが順に現れている．したがって，kが小さい方から順に検定をおこなった場合，$k=1$では5 mm付近のピークに加えて3.0～3.5 mm付近の小さなピークを認めるのか，$k=2$ではさらにそれ以外にもピークを認めるべきなのかを，それぞれ問題にしていることになる．

$k=1$で検定すると，ブートストラップ回数10,000回で$\hat{P}^1_{\mathrm{boot}}$は0.0165となり（表1），有意水準5％で帰無仮説は棄却された．一方$k=2$での検定では，$\hat{P}^2_{\mathrm{boot}}=0.4505$となり（表1），帰無仮説$\mathrm{H}_0^2$は棄却されない．したがって，この検定結果からは，下顎類歯の類舌径の計測値は2つの部分母集団に由来していると考えて良さそうである．

2．ポリプチコセラスの最終隔壁

最終隔壁の位置データから算出された臨界バンド幅$h_{1.crit}$，$h_{2.crit}$，$h_{3.crit}$はそれぞれ2.690038，1.952924，1.599452であった（表1）．$h_{1.crit}$のときは第3シャフト中盤にある最も大きなピーク[β]付近に極大をもち，$h_{2.crit}$のときはそれに加えて第2シャフト終盤にあるピーク[α]が極大として現れ，$h_{3.crit}$ではさらに第4シャフトのピーク[γ]が極大として加わる（図9）．したがってこの例でも，いずれもモードとして期待されていたものが順に現れていることがわかる．kが小さい方から順に検定をおこなえば，$k=1$ではピーク[β]に加えてピーク[α]を認めるのか，$k=2$ではピーク[γ]を認めるのか，そして$k=3$ではそれ以外にもピークを認めるべきなのかを，それぞれ問題にしていることになる．

$k=1$から順に検定すると，ブートストラップ回数10,000回で$\hat{P}^1_{\mathrm{boot}}$，$\hat{P}^2_{\mathrm{boot}}$，$\hat{P}^3_{\mathrm{boot}}$はそれぞれ0.0102，0.0290，0.0110となり（表1），いずれも有意水準5％で帰無仮説は棄却された．$k=3$での検定は，最終隔壁の位置の分布が三峰性であるという帰無仮説の検定であるから，本来H_0^3は棄却されないことが期待されたのだが，Silvermanの検定では，二峰性よりも三峰性の方がより“ありそうにない”と判断されたらしい．

試しに$k=4$のカーネル密度推定をおこなったところ，$h_{4.crit}=0.872359$であり（表1），そのときピーク[α]とピーク[β]との間に小さな極大[δ]を認めている．検定では，$\hat{P}^4_{\mathrm{boot}}=0.6126$となり（表1）$\mathrm{H}_0^4$は棄却されない．すなわちSilvermanの検定の結果のみに基づけば，最終隔壁の位置は四峰性分布であるという結論になる．

しかしこの検定結果はSilvermanの検定の特徴をよく表している．Silvermanの検定では，計量に複数のピークがあった場合，他とより離れているピークがモードとして認識され易い．他と離れたモードを消すためには，より大きなバンド幅が必要となるからである．今回のデータでは，ピーク[α]は見た目には目立たない小さなピークであるにもかかわらず，大きなピーク[β]と位置が離れていたため，ピーク[γ]よりも先にモードとして認識されている．同様に，4つ目の極大[δ]もまた，他のピークとやや離れたところにデータがあったため，極大として現れやすかったと考えられる．一方でピーク[γ]はそ

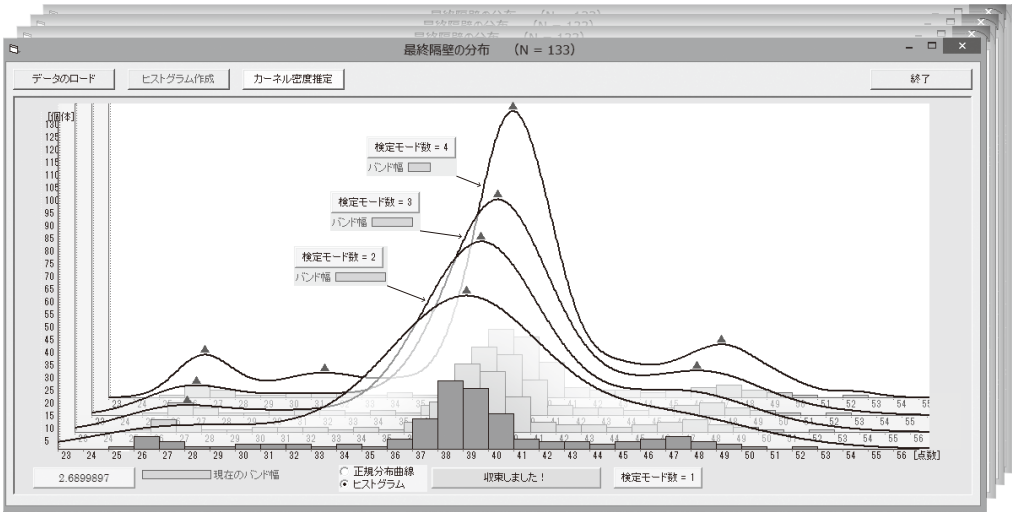


図9．自作プログラムVISUAL-SILVERMANによって流矢層産ポリプチコセラスの最終隔壁の分布を解析した例．帰無仮説が主張する確率密度分布を求めた画面（図8bに相当）を$k=1$から$k=4$までについて合成して示した．続いて行われるSilvermanの検定では，それぞれ，新たに増えんとするピークの認否が焦点になる．

こそこ大きく見た目には明瞭に見えるにもかかわらず，大きなピーク[β]とあまり離れていないために極大として現れにくい状況になっていた．そのため，ピーク[γ]を極大とするならば，第2ターンにも極大[δ]があると考えられる，という検定結果になったものであろう．

最後に現れた小さな極大[δ]は，実際には，サンプリング・バイアスによるものかもしれない．この例では数多くの部分的な標本からもデータを取っている．そういった場合，ターン部からの距離によって最終隔壁の位置を決めたのだが，両側共にターン部が残っていない棒状の標本は無視せざるを得なかったのである．必然，ターン部の近傍では観測機会が密になり，シャフトの中央部付近では疎になるというサンプリング・バイアスが掛ってしまった．特に観測機会が密な第2ターン付近ではピーク[α]の右側の裾にこのバイアスが相乗されて見かけの極大[δ]が現われた可能性がある．

著者の一人岡本はこの目立たないピークに関する上述の問題に気づいていたのだが，数値的に検出されることはあるまいと考えて当時は放置してしまった．そのためSilvermanの検定によって，このような微小なピークまで指摘されたことには正直驚かされた．この検定に実際そこまでの切れ味があるのか，あるいはまぐれ当たりだったのかは，なお検討する余地があるだろう．しかし，少なくとも古生物の研究者が手にした分布型についての意味を考える際に，この検定が非常に有用である点に疑いはないと考える．

Silvermanの検定の問題点

Silvermanの検定に対しては，いくつかの批判があることも事実である．代表的な批判としては，(1) 検定結果が保守的になり帰無仮説が棄却されにくい，つまり検定自体の性質として第二種過誤の起こる確率が高い（例えばSilverman, 1983; Mammen *et al.*, 1992; Hall and York, 2001），(2) $f_0^k(x)$の分散調整が妥当か（例えばFisher *et al.*, 1994），(3) 分布の裾の方で偽のモードが出てしまう（例えばIzenman and Sommer, 1988; Fisher *et al.*, 1994; Hall and York, 2001），といったことが挙げられる．

保守的な検定結果

検定結果が保守的になるのは，Silvermanの検定では$\Pr_{H_0^k}\{\hat{h}^*_{k.crit} > \hat{h}_{k.crit}\}$の分布が，少なくとも$n$が大きければ，(0, 1) 区間でほぼ一様になることを前提にしているが，実際にはそうではないためらしい（Hall and York, 2001）．つまりSilvermanの検定のP値を

$$P^k_{\mathrm{boot}} = \Pr_{H_0^k}\{\hat{h}^*_{k.crit} > \lambda_\alpha \hat{h}_{k.crit}\}$$

と書けば，有意水準αでのSilvermanの検定では$\lambda_\alpha \equiv 1$として$\hat{P}^k_{\mathrm{boot}} < \alpha$で$H_0^k$を棄却するが，実際には$\lambda_\alpha > 1$なのである．Hall and York（2001）によれば$k=1$での検定で得られたP値が0.05のとき，実際の水準は0.010である（Hall and York, 2001, table1）．上の適用例で言えば，ポリプチコセラスに関する検定で$k=1$のときの実際のP値はほぼ0らしい．彼らはさらに$k=1$の場合，

$$\lambda_\alpha = \frac{a_1\alpha^3 + a_2\alpha^2 + a_3\alpha + a_4}{\alpha^3 + a_5\alpha^2 + a_6\alpha + a_7}$$

の形で表せば$a_1 = 0.94029$，$a_2 = -1.59914$，$a_3 = 0.17695$，$a_4 = 0.48971$，$a_5 = -1.77793$，$a_6 = 0.36162$，$a_7 = 0.42423$であることを示している．したがって，$k=1$での検定にはHall and York（2001）のキャリブレーションを利用すれば，検定結果が保守的になる問題は解決できる．しかし$k \geq 2$についてのこの種のキャリブレーションは，現時点

でおそらく提案されていない．

分散の調整

次に$(1+h_{k.crit}^2/s^2)^{-1/2}$による$f_0^k(x)$の分散の調整であるが，$k=1$のときの$f_0^1(x)$に対してであれば良いが，$k>1$ではあまり説得力がないとされる（Fisher *et al.*, 1994）．$f_0^k(x)$の分散を調整することは，すなわちバンド幅を調整していることになる．そこでFisher *et al.*（1994）はブートストラップ抽出の際に分散の調整をおこなうのではなく，P値を，

$$P_{\mathrm{boot}}^k=\Pr_{\mathrm{H}_0^k}\{h_{k.crit}^*>Rh_{k.crit}\}$$

とし，Rによって$h_{k.crit}$を$h_{k.crit}^*$と同程度の平滑化スケールに調整することを提案している．Rの推定法についてはFisher *et al.*（1994）を参照されたい．

しかしこの方法を適用したFisher *et al.*（1994）のシミュレーション結果によれば，検定結果にはさほど大きな差が出るとは言えないようである．単峰性分布や顕著な二峰性分布ではFisherらの方法は効果的なようだが，それほど顕著でない二峰性分布になると，Fisherらの方法とSilvermanの方法とであまり差は見られない．その後の研究でもこの問題に関する効果的な解決方法は報告されていないようである．

偽のモード

分布の裾の方で偽のモードが出てしまう点については，カーネル密度推定自体の問題である．裾の長い分布からデータが得られる場合，分布の裾の方のデータはまばらになる．したがって，そのようなデータに関して，バンド幅を小さくとって密度推定をおこなったとき，$\hat{f}_h(x)$はデータのまばらな裾の方にも小さなモードの存在が推定されてしまうことがある（図10）．その場合，Silvermanの検定ではモード数が実際よりも多く出てしまう可能性がある．上の適用例のうちポリプチコセラスに関する検定で現れた第4の極大[δ]はこのことと関連していると考えられる．

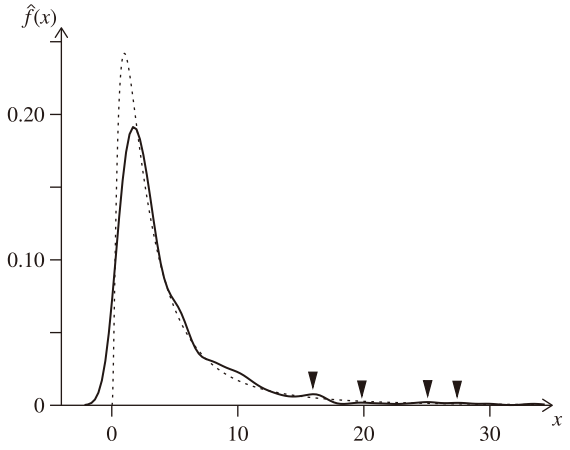


図10．対数正規分布LN(1, 1)の曲線（点線）とそこから無作為抽出された$n=500$の標本に基づいて見積もられたカーネル密度推定量（実線）．標本から算出された適正なバンド幅（$h=0.778$; Silverman, 1986）で密度推定をおこなうと，再構成された分布の裾に複数の偽モードが現れてしまう（黒矢印）．

カーネル密度推定のこの問題は，データが密な部分と疎な部分とで同じバンド幅を使って密度推定をおこなうために生じる．それならば，バンド幅を可変にすれば良い．そこで考案されたのがadaptiveカーネル密度推定である（Abramson, 1982a, b）．一次のadaptiveカーネル密度推定量は，

$$\hat{f}_h(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\mu_i}K\left(\frac{x-x_i}{h\mu_i}\right)$$

で与えられる（Silverman, 1986; Izenman, 1991）．ただしμ_iはlocal bandwidth factorsである．Izenman and Sommer（1988）は，このadaptiveカーネル密度推定量をSilvermanの検定に用いることを提案している．

古生物学での利用

古生物学の研究でSilvermanの検定を用いる場合，上のような問題点に関しても知った上で使う必要があろう．結果が保守的になる点については，$k=1$の場合であればHall and York（2001）のキャリブレーションを使えば良い．また分散の調整についても$k=1$であれば問題にはならない．したがって，現時点ではモードの数を知るというよりも，$k\geq 2$であることを検定するために利用するのが最も無難だと考えられる．その場合，Hall and York（2001）のキャリブレーションを使わなくてもH_0^1を棄却できるのであれば，必ずしも使う必要はない．

裾の方で偽のモードが出る問題については，慎重な判断が必要となる．古生物学でこの検定をおこなう場合，明らかに他から飛び離れたデータがあれば，まずはそのデータを吟味すべきだろう．例えば計測ミスや他と同等に比較できないデータであれば，一緒にして検定をおこなう意味はないからである．

検定の際にadaptiveカーネル密度推定を用いる方法では，本来意味のある小さなモードが検出されないおそれがある．それよりも，実際に現れたモードの位置を確認した上で，その意味を検討する方が良いように思われる．例えばH_0^1が棄却されたならば，$h_{2.crit}$を使ったときの$\hat{f}_{h_{2.crit}}(x)$のグラフを見れば，次に出現するモードの位置を確認できる．またCheng and Hall（1998）は，高さが$K(0)/nh$の1.5倍（$3/2nh\sqrt{2\pi}$）以上のモードのみに限って数えれば良いと述べている．いずれにせよ検定によって大きなモードの裾の方にある小さなモードに意味を持たせたい場合には，検定結果のみに頼らず，小さなモードの存在を主張する根拠を他にも示すべきだろう．

Silvermanの検定のみに限らず，検定はあくまでもデータを評価するための一手段に過ぎない．最も重要なこと

は，得られたデータを古生物学の研究に携わる者として如何に解釈するかであり，検定結果はそのための判断材料の1つと位置付けるべきである．古生物学で扱うデータは，統計学的手法のみに依拠して結論付けられるようなものでは決してない．我々に必要な「統計力」とは，統計学的手法を正しく理解した上で，我々ならではの研究にそれを利用する力であろう．得られたデータの個性をできる限り生かしつつ，それが示す意味をより客観的な視点で読み取るのが我々の責務であろうし，そのために利用すればこそ，統計学は有用な手法たり得るのではないだろうか．

謝辞

本稿は成瀬　元氏（京都大学）らお二人の査読者と編集者の前田晴良氏（九州大学）のご指摘により大きく改善された．記して感謝の意を表する．

文献

Abramson, I.S., 1982a. On bandwidth variation in kernel estimates –a square root law. *Annals of Statistics*, **10**, 1217–1223.

Abramson, I.S., 1982b. Arbitrariness of the pilot estimator in adaptive kernel methods. *Journal of Multivariate analysis*, **12**, 562–567.

Ahmed, M.O. and Walther G., 2012. Investigating the multimodality of multivariate data with principal curves. *Computational Statistics and Data Analysis*, **56**, 4462–4469.

Aitkin, M. and Rubin, D.B., 1985. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society B*, **47**, 67–75.

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. 482p., Clarendon Press, Oxford.

Brandon, M.T., 1992. Decomposition of fission-track grain-age distributions. *American Journal of Science*, **292**, 535–564.

Burman, P. and Polonik, W., 2009. Multivariate mode hunting: Data analytic tools with measures of significance. *Journal of Multivariate Analysis*, **100**, 1198–1218.

Cacoullos, T., 1966. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, **18**, 179–189.

Chan, K.S. and Tong H., 2004. Testing for multimodality with dependent data. *Biometrika*, **91**, 113–123.

Chaudhuri, P. and Marron, J.S., 1999. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–823.

Cheng, M.-Y. and Hall, P., 1998. Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society B*, **60**, 579–589.

Davison, A.C. and Hinkley, D.V., 1997. *Bootstrap Methods and their Application*. 582p., Cambridge University Press, Cambridge.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

Efron, B. and Tibshirani, R.J., 1994. *An Introduction to the Bootstrap*. 436p., Chapman and Hall/CRC, Boca Raton.

Epanechnikov, V.K., 1969. Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, **14**, 153–158.

Everitt, B.S., 1996. An introduction to finite mixture distributions. *Statistical Methods in Medical Research*, **5**, 107–127.

Fisher, N.I., 1989. Smoothing a sample of circular data. *Journal of Structural Geology*, **11**, 775–778.

Fisher, N.I., Mammen, E. and Marron, J.S., 1994. Testing for multimodality. *Computational Statistis and Data Analysis*, **18**, 499–512.

Good, I.J., and Gaskins, R.A., 1980. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, **75**, 42–56.

Hall, P., Minnotte, M.C. and Zhang, C.-M., 2004. Bunp hunting with non-Gaussian kernels. *Annals of Statistics*, **32**, 2124–2141.

Hall, P. and York, M., 2001. On the calibration of Silverman's test for multimodality. *Statistica Sinica*, **11**, 515–536.

Hartigan, J.A. and Hartigan, P.M., 1985. The dip test of unimodality. *Annals of Statistics*, **13**, 70–84.

Hartigan, J.A. and Mohanty, S., 1992. The RUNT test for multimodality. *Journal of Classification*, **9**, 63–70.

Hartigan, P.M., 1985. Computation of the dip statistic to test for unimodality. *Applied Statistics*, **34**, 320–325.

速水　格，1969．化石の計測と統計—古生物学実習の1例—．九州大学理学部研究報告（地質学），**10**，67–90.

速水　格・松隈明彦，1971．化石の計測と統計—アロメトリーと個体変異の解析—．九州大学理学部研究報告（地質学），**10**，135–160.

Ihaka, R. and Gentleman, R., 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Izenman, A.J., 1991. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, **86**, 205–224.

Izenman, A.J. and Sommer, C.J., 1988. Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, **83**, 941–953.

Johnson, D.H., 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management*, **63**, 763–772.

Jones, M.C., 1991. On correcting for variance inflation in kernel density estimation. *Computational Statistics and Data Analysis*, **11**, 3–15.

金田尚久・新居玄武，2009a．混合分布問題—その基礎からカーネル降下法まで—Part 1．学習院大学経済論集，**46**，1–30.

金田尚久・新居玄武，2009b．混合分布問題—その基礎からカーネル降下法まで—Part 2．学習院大学経済論集，**46**，127–170.

Lo, Y.-T., Mendell, N.R. and Rubin, D.B., 2001. Testing the number of components in a normal mixture. *Biometrika*, **88**, 767–778.

Mammen, E., Marron, J.S. and Fisher, N.I., 1992. Some asymptotics for multimodality tests based on kernel density estimates. *Probability Theory and Related Fields*, **91**, 115–132.

松岡廣繁，2000，石川県白峰村桑島の“化石壁”から産出したトリティロドン科哺乳類型爬虫類化石について．松岡廣繁編，石川県白峰村桑島化石壁の古生物—下部白亜系手取層群桑島層の化石群—，53–74．石川県白峰村教育委員会.

McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society C*, **36**, 318–324.

McLachlan, G.J. and Peel, D., 2000. *Finite Mixture Models*. 419p., John Wiley & Sons, Inc., New York.

Minnotte, M.C., Marchette, D.J. and Wegman, E.J., 1998. The bumpy road to the mode forest. *Journal of Computational and Graphical Statistics*, **7**, 239–251.

Minnotte, M.C. and Scott, D.W., 1993. The mode tree: a tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics*, **2**, 51–68.

Müller, D.W. and Sawitzki, G., 1991a. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, **86**, 738–746.

Müller, D.W. and Sawitzki, G., 1991b. Using excess mass estimates to investigate the modality of a distribution. *In* Öztürk, A., van der Meulen, E.C., Dudewicz, E.J., and Nelsen, P.R., *eds.*, *The Frontiers of Statistical Scientific Theory and Industrial Applications, vol. II*, 355–382. American Sciences Press, Syracuse.

Okamoto, T. and Shibata, M., 1997. A cyclic mode of shell growth and its implications in a Late Cretaceous heteromorph ammonite *Polyptychoceras pseudogaultinum* (Yokoyama). *Paleontological Research*, **1**, 29–46.

Parzen, E., 1962. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065–1076.

Pearson, K., 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, **185**, 71–110.

Polonik, W., 1995. Measuring mass concentrations and estimating density contour clusters—An excess mass approach. *Annals of Statistics*, **23**, 855–881.

R Core Team, 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at www.r-project.org.

Render R.A. and Walker, H.F., 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195–239.

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832–837.

Rozál, G.P.M. and Hartigan, J.A., 1994. The MAP test for multimodality. *Journal of Classification*, **11**, 5–36.

Schwaiger, F. and Holzmann, H., 2013. Silvermantest: package which implements the silvermantest. Available at www.uni-marburg.de/fb12/stoch/.

Silverman, B.W., 1981. Using kernel density estimates to investigate multimodality. *Journal of Royal Statistical Society B*, **43**, 97–99.

Silverman, B.W., 1983. Some properties of a test for multimodality based on kernel density estimates. *In* Kingman, J.F.C. and Reuter, G.E.H., *eds.*, *Probability, Statistics and Analysis, London Mathematical Society Lecture Note Series 79*, 248–259. Cambridge University Press, London.

Silverman, B.W., 1986. *Density Estimation for Statistics and Data Analysis*. 175p., Chapman and Hall, London.

樋田　勉，1999．多峰性検定とプラグイン・バンド幅決定法．早稲田経済学研究，(48)，83–95.

汪　金芳・桜井裕仁，2011．Rで学ぶデータサイエンス，**4**，ブートストラップ入門．236p.，共立出版.

Wong, M.A., 1985. A bootstrap testing procedure for investigating the number of subpopulations. *Journal of Statistical Computation and Simulation*, **22**, 99–112.

（2014年3月18日受付，2014年9月12日受理）