統計数理（1996）
第44巻 第1号 3-18

特集「計算統計学の発展」
[原著論文]

ブートストラップ法 —— 2標本問題からの考察

統計数理研究所　汪　金芳
千葉大学*　田栗正章

（1995年5月　受付）

概　要

ブートストラップ法は，1979年頃B. Efronによって提唱された方法で，計算機を活用してデータの持つ情報を抽出しようとするものである．それは，観測されたデータから，計算機を用いてリサンプリングを行うことにより，1組の標本だけからでは解明不可能な，または解明することが困難な統計的諸性質についての検討を行おうとするものである．そのような性質としては，推定量のバイアス，分散，予測誤差や，より複雑な信頼区間，分布関数のようなものが対象となっている．ブートストラップ法は，これらの性質についての評価を，計算機アルゴリズムとして記述可能な自動的な形式で行うものである．

本論文では，これまでにも様々な解析の対象とされてきた，ダーウィンのデータに基づく2標本問題の例を用いて，ブートストラップ法を紹介する．それはブートストラップ法を，分かり易く，実際問題に即した形で，かつ最新の成果に至るまで解説することができると考えたからである．いくつかの例を通して，統計的データ解析を行う際に，ブートストラップ法が簡単・明瞭で，多種多様な場合に適用可能で，かつ極めて有効な方法であることが分かるだろう．そしてブートストラップ法は，実際問題の解析に携わる統計家にとって，なくてはならぬ手法の1つであることが納得できるであろう．

本論文ではまた，ブートストラップ検定に関する新しい方法についても議論を行う．すなわち等分散性の仮定のもとで2つの母平均が等しいか否かを検定するための，混合ブートストラップ検定の方法を提案する．またリサンプリング推定方程式という観点から，ブートストラップ法の現在の発展状況と今後の方向についても考察を行う．

1. 統計学，計算機，そしてブートストラップ法

統計学の目的の1つは，例えば表1に示されているようなデータに対してその縮約を行い，情報の抽出を行うことである．表1のデータはDarwin（1876）によって与えられたものであり，自家授精に対する他家授精の優越性の有無を検証するために採られたデータである．ダーウィンは7種類の植物についての実験を行ったが，ここではそれらの内「とうもろこし」のデータのみを表に示している．いまμ_x, μ_yを，それぞれ他家授精，自家授精したとうもろこしの丈の高さの平均とすれば，この場合の統計的な問題は，帰無仮説$H_0: \mu_x = \mu_y$に対して対立仮説$H_1: \mu_x > \mu_y$を検定することである．

情報抽出を目的とするデータの縮約に関して，Fisher（(1922), p. 313）は，統計的推測の問題を次の3種類に分類している．すなわち(i)母集団分布想定の問題，(ii)推定，検定の問題，(iii)標

* 理学部：〒263 千葉市稲毛区弥生町1-33.

表1．他家授精,自家授精したとうもろこしの丈に関するダーウィンの観測値（1/8インチ単位).

他家受精	188	96	168	176	153	172	177	163
(x)	146	173	186	168	177	184	96	
自家受精	139	163	160	160	147	149	149	122
(y)	132	144	130	144	102	124	144	

註：このデータの出典は Fisher ((1960), p. 30) であるが，ここでは簡単のために植木鉢の情報は無視してある.

本分布決定の問題である．Fisher (1922) は，第1の問題は実務に携わる統計家が決定すべきものである，としている．しかし量子力学における不確定性原理と基礎方程式が古くからあるように，母集団分布（族）も統計的不確定性関係に基づく統計基礎方程式を解くことによって得られる，という考え方もある（松縄 (1994)）．また母集団分布（族）を特定しなくても，2次までのモーメントしか仮定しない，擬似尤度に基づく推測 (quasi-likelihood inference) も可能である (Wedderburn (1974), McCullagh and Nelder (1989), Chapter 9, Fahrmeir and Tutz (1994) など参照)．ここで取り扱うダーウィンの例では，通常2つのデータの組は，正規母集団 $N(\mu_x, \sigma_x^2)$ および $N(\mu_y, \sigma_y^2)$ から互いに独立に取られたことを仮定している.

次に(ii), (iii)の問題を，検定の場合について同時に考えてみよう．すなわちこの場合の問題は，例えば次式 (1.1) で与えられる通常の t 統計量のような検定統計量の選択の問題と，その帰無仮説 $H_0: \mu_x = \mu_y$ のもとでの標本分布の決定の問題である．ただし $s_x^2 = \sum (x_i - \bar{x})^2/(m-1)$, $s_y^2 = \sum (y_i - \bar{y})^2/(n-1)$ である.

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/m + s_y^2/n}} \tag{1.1}$$

Fisher ((1960), p. 37) は，x, y に関する2組のデータが対になっているものと考え，それらの差を $188-139=49$, $96-163=-67$, ..., $96-144=-48$ などと計算し，正規性の仮定のもとで1標本 t 検定を行った．フィッシャーは，同じデータに対して行った並べかえ検定の結果を引用して，彼の行った t 検定が妥当であることを主張した.

ところで，フィッシャーの解析でまず問題となるのが正規性の仮定である．そしてさらに問題なのがデータの対を作って解析していることである．表1の上段と下段のデータはもともと対をなしている訳ではないから，本当のノンパラメトリックな並べかえ検定では，30!/15!15! = 155,117,520 個の平均の差を対象としなければならないはずである．しかしこの代わりにフィッシャーは，$2^{15} = 32{,}768$ 個の差だけを対象にしている．これは，計算機のなかったフィッシャーの時代に真の並べかえ検定を行おうとすれば，表1のような小さなデータの組に対してでさえ何年もの計算が必要になり，とても対処しきれなかったからであろう.

さてブートストラップ法によれば，(1.1) で与えた統計量の帰無仮説のもとでの分布を，計算機を用いて自動的に近似計算することができる．この場合モデルの仮定はまったく必要ではなく，計算は与えられたデータだけに基づいて行われるのである．ブートストラップ法は，いかなる状況下でも解析手順が自動化されているという意味において，並べかえ検定より一般的な方法であるということができよう．それは観測されたデータのみに基づいて，様々な統計的諸性質を評価しようとする極めて一般的な方法である．その対象は，推定量のバイアス，分散，予測誤差や，より複雑な信頼区間，分布関数などである．第2～4節で述べるように，ブートストラップ法によれば，これらすべての評価が，計算機を用いて自動的に行えるのである．仮説検定などの場合に

は，戦後の統計の世界を席巻したネイマン-ピアソン流の枠組みに基づく決定理論的なアプローチが依然として根強いように思える．しかし推定の一般論の場においては，フィッシャーの統計的推論に対する考え方が表舞台に再登場してきたと考えられる（Efron and Tibshirani (1993), p. 393, Reid (1994) 参照）．フィッシャー流の言い方をすれば，ブートストラップ法とは，複雑な統計的問題に対して，ノンパラメトリックな場合の尤度最大化原理を適用する方法と言えるかもしれない．そしてこれは，計算機の強力なパワーを用いて自動的に行われるのである．

簡潔に言えば，ブートストラップ法は，上述した(iii)のタイプの問題を取り扱う，データ依存型のシミュレーション法である．広い範囲の問題に対するブートストラップ法の妥当性を初めて示したのは，Singh (1981), Bickel and Freedman (1981), Freedman (1981) などである．また Hall ((1992), Chapter 3) は，多くの場合において伝統的な正規理論に基づく推論と比較して，より正確な情報を提供する方法であると言っている．ブートストラップ法の手順の決定には，漸近展開のような複雑な計算は必要でない．それは簡単・明瞭で，様々な場合に適用可能で，かつ極めて有効な統計的データ解析の方法である．そしてブートストラップ法は，実際問題の解析に携わる統計家にとって，なくてはならぬ手法の1つであろう．

次の第2節では，比較的簡単な統計的誤差の推定の場合について，ブートストラップ法とそれに深く関連するジャックナイフ法の考察を行う．第3節および第4節では，ブートストラップ法による信頼区間の構成と，より複雑な2標本問題に対する仮説検定の方法について議論を行う．統計的誤差の推定や信頼区間の構成の場合には，ブートストラップ法は現在では確立した方法となっているが，仮説検定の場合には数多くの検討すべき問題が残されている．本論文の最後の節では，ブートストラップ法の現在の発展状況と今後の方向について概観しておく．

2. 種々の統計的誤差のブートストラップ推定

ブートストラップ法は，歴史的にはバイアス，分散，予測誤差などのような統計的誤差の推定を行うノンパラメトリックな方法として提案されたもので（Efron (1979)），それまでにある程度確立されていたジャックナイフ法（Quenouille (1956)）の対案であった．母集団分布 F から取られた大きさ n の *i.i.d.* 標本 $z_1,\ldots,z_n$ に基づく，ある推定量 $\hat{\theta}(z_1,\ldots,z_n)$ のバイアスや分散を推定するのは，一般にそう易しい問題ではない．次の比推定量 $\hat{r}$ と相関係数推定量 $\hat{\rho}$ のような単純でない形をしている推定量の場合を考えてみれば分かるであろう．

$$\hat{r}=\bar{x}/\bar{y},\qquad \hat{\rho}=\frac{\overline{xy}-\bar{x}\bar{y}}{\sqrt{\overline{x^2}-\bar{x}^2}\sqrt{\overline{y^2}-\bar{y}^2}}. \tag{2.1}$$

ブートストラップ法やジャックナイフ法によれば，$\hat{\theta}$ の複雑性に関係なくバイアスや分散などの統計的誤差の推定が自動的に行える．例えばジャックナイフ法による $\hat{\theta}$ のバイアスおよび分散推定量は，それぞれ次式のように簡単な形で与えられる．

$$\hat{b}_J=(n-1)(\hat{\theta}_{(\cdot)}-\hat{\theta}), \tag{2.2}$$

$$\hat{\sigma}_J^2=(n-1)\sum(\hat{\theta}_{(i)}-\hat{\theta}_{(\cdot)})^2/n, \tag{2.3}$$

ここで $\hat{\theta}_{(\cdot)}=\sum\hat{\theta}_{(i)}/n$ は，i 番目の観測値を除外した標本から得られる推定量 $\hat{\theta}_{(i)}$ の平均である．このような推定量が論拠を持つためには，ある種の正則条件が必要である．例えば $\hat{\theta}$ が分布の滑らかな汎関数統計量になっている，すなわちある滑らかな汎関数 $\theta(\cdot)$ に対して $\hat{\theta}=\theta(F_n)$ と書けることなどが仮定されている．したがって中央値やトリム平均などは，この推定量のクラスには入らない．このような場合には，分散の推定に関しては，ブートストラップ法の方がより信

頼できる（Efron (1982), p. 28).

ブートストラップ法は，考え方の上ではジャックナイフ法より簡単である．推定量$\hat{\theta}$のバイアスや分散は一般に分布Fに依存するから，それを$b(F)$, $\sigma^2(F)$と書くことにすれば，ブートストラップ法は単にFをF_nで置き換えて推定を行うという方法である．ここでF_nは，n個の観測値$z_1,\ldots, z_n$の各点に$1/n$のマスを与えた経験分布関数である．一般には$b(F_n)$, $\sigma^2(F_n)$は解析的に求められないので，以下のアルゴリズムに示すように，それは通常モンテカルロ近似で置き換えられる．

1．母集団分布Fから取られた1組の観測値$\boldsymbol{z}=\{z_1,\ldots, z_n\}$から，大きさ$n$の無作為標本$\boldsymbol{z}_b^*=\{z_{b1}^*,\ldots, z_{bn}^*\}$を復元抽出する．
2．$\boldsymbol{z}$の代わりに$\boldsymbol{z}_b^*$を用いて$\hat{\theta}$を計算し，それを$\hat{\theta}_b^*=\hat{\theta}(z_{b1}^*,\ldots, z_{bn}^*)$とおく．
3．ステップ1と2をB回繰り返し，$\hat{\theta}_1^*,\ldots,\hat{\theta}_B^*$を計算する．
4．$b(F_n)$, $\sigma^2(F_n)$に対するモンテカルロ近似値を

$$\hat{b}=\hat{\theta}_{(\cdot)}^*-\hat{\theta},\qquad \hat{\sigma}^2=\sum_{b=1}^{B}(\hat{\theta}_b^*-\hat{\theta}_{(\cdot)}^*)^2/(B-1) \tag{2.4}$$

で与える．ここで，$\hat{\theta}_{(\cdot)}^*=\sum_{b=1}^{B}\hat{\theta}_b^*/B$である．

実際に上のアルゴリズムを遂行する際には，シミュレーション回数Bは通常50から200程度で十分である（Efron (1987)）．計算機環境が飛躍的に発達している今日においては，この程度の計算量は問題とはならない．

上のアルゴリズムは極めて一般的である．すなわち（2.1）で与えられた比推定量$\hat{r}$や相関係数推定量$\hat{\rho}$などの場合のように，$z_1,\ldots, z_n$がベクトル観測値であっても，その計算手順はまったく同じである．ここで表1に与えたダーウィンのデータを用いて，$\hat{r}$と$\hat{\rho}$に対するバイアスおよび分散推定量の値を計算してみよう．この場合与えられる観測値は2変量でなければならないので，ここではフィッシャーが行ったのと同じデータの対を作成して計算を行うことにする．すなわち$z_1=(188, 139)$, $z_2=(96, 163),\ldots, z_{15}=(96, 144)$と考えることにする．下の表2は，上記2種類の推定量に対するバイアスおよび分散推定値の計算結果をまとめたものである．（比推定や相関係数の推定の場合には，上で定義したバイアス及び分散推定量は不偏ではないので，リサンプリングの手順を2重，3重に繰り返すことによって，そのバイアスを高次のオーダーまで修正することは可能である．しかしそのような場合のアルゴリズムや解析は複雑になるので，ここでは省略する．）ここでブートストラップ近似値は2000組のブートストラップ標本に基づくものである．また表2に与えたデルタ法による近似は，母集団分布$F(x, y)$に正規性を仮定したときの漸近バイアスと漸近分散の推定値である．例えば，$\hat{\rho}$の分散のデルタ法による推定値は，

表2．ダーウィンのデータの場合の，比推定量$\hat{r}$と相関係数推定量$\hat{\rho}$に対するバイアスおよび分散の推定値．

方法	$\hat{r}$		$\hat{\rho}$	
	偏り	分散	偏り	分散
ジャックナイフ	.0016	.0052	.0198	.050
ブートストラップ	.0010	.0050	.0085	.038
デルタ法	.0030	.0053	.0099	.066

註：デルタ法と記した場合の値は，正規分布を仮定したときの推定量の漸近バイアスと漸近分散の推定値である．

$(1-\hat{\rho}^2)^2/(n-3) = (1-(-0.3348)^2)^2/12 \approx 0.066$ (Johnson and Kotz (1970), p. 229参照)と計算できる．デルタ法は本質的には線形近似法であるが，ブートストラップ法は推定量のより高次の曲がり具合いまで考慮に入れている．この例においては，ブートストラップ法によるバイアスおよび分散の推定値はジャックナイフ法のそれぞれの推定値より小さくなっている．この現象はある種の条件の下で正確に証明することができる (Efron (1982), Chapter 6).

3. ブートストラップ信頼区間

母集団分布 F の滑らかな汎関数として与えられるパラメータ $\theta = \theta(F)$ に対する信頼区間の構成を考えてみよう．伝統的には，θ に対する推定量 $\hat{\theta} = \theta(F_n)$ またはそれをスチューデント化した推定量の標本分布を近似的に求め，それから信頼限界を計算するのが一般的な方法であった．これを精密に実行する1つの方法は漸近展開であるが，その場合には F の関数型は既知としなければならない．

これに対してブートストラップ法では，分布の型を仮定する必要はもちろんなく，また例えばエッジワース展開のような数学的に複雑な操作を行う必要もない．ブートストラップ信頼区間は，与えられるデータと推定量の形だけから，次のようなアルゴリズムによって簡単に構成できる．

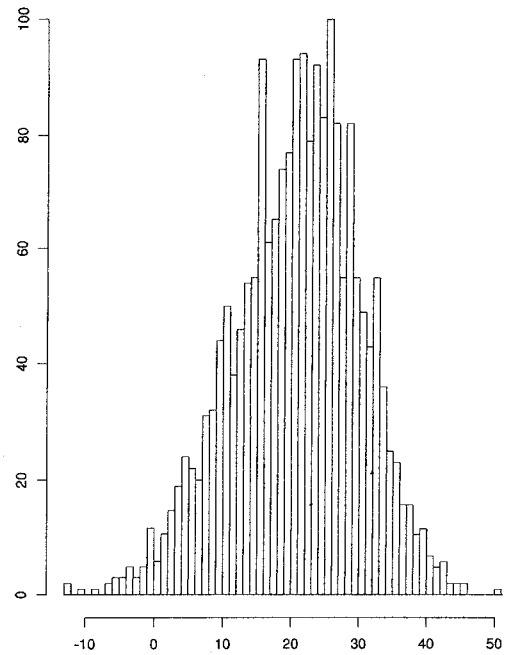
1．母集団分布 F から取られた1組の観測値 $\boldsymbol{z} = \{z_1, \ldots, z_n\}$ から，大きさ n の無作為標本 $\boldsymbol{z}_b^* = \{z_{b1}^*, \ldots, z_{bn}^*\}$ を復元抽出する．
2．$\boldsymbol{z}$ の代わりに $\boldsymbol{z}_b^*$ を用いて $\hat{\theta}$ を計算し，それを $\hat{\theta}_b^* = \hat{\theta}(z_{b1}^*, \ldots, z_{bn}^*)$ とおく．
3．ステップ1と2を B 回繰り返し，$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ を計算する．
4．$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ を大きさの順に並べ変える．
5．パラメータ θ に対する信頼係数 $1-2\alpha$ の信頼区間を

$$[\hat{\theta}_{low}^*,\ \hat{\theta}_{up}^*] \tag{3.1}$$

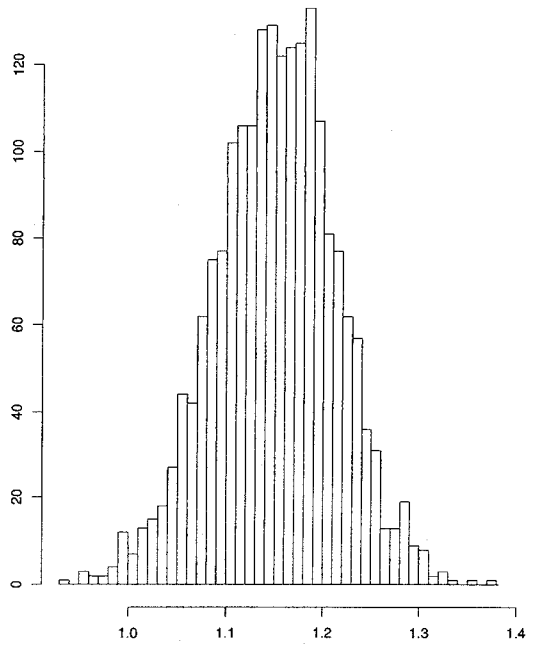
で構成する．ここで low と up は，それぞれ $\alpha = low \times B$, $1-\alpha = up \times B$ を満たす整数である．(この式を満たす low や up の整数値が存在しない場合には，B を調整するか，または最も近い整数値を用いて近似的に信頼係数 $1-2\alpha$ の信頼区間を構成するかを選択するのが普通である．)

(3.1) はパーセンタイル・ブートストラップ信頼区間と呼ばれている．この方法は極めて単純であるが，その導出過程では理論的に多少粗い部分がある．そのため多くの研究が行われ，これを精密化するいくつかの方法が提案されている．それらの中で有力な方法に，abc 法，bc_a 法，ブートストラップ $-t$ 法などがある．これらの方法とその理論的な性質については，Hall ((1992), Chapter 3), Efron and Tibshirani ((1993), Chapter 12-14) に詳説されている．また漸近的に bc_a 法と同値な方法である，自動パーセンタイル法 (DiCiccio and Romano (1995)) も注目すべきであろう．

このようにブートストラップ法によれば，前節の統計的誤差の推定の場合と同様，$\hat{\theta}$ の標本分布を，例えば図1に示したようなモンテカルロ法を適用して作られるブートストラップ・ヒストグラムによって，いつでも容易に近似できる．いま，表1のとうもろこしのデータに関して，2つの母集団平均の差 $\theta = \mu_x - \mu_y$ に対する信頼係数95%の信頼区間の構成を考えてみよう．この場合初期標本を，表1において対応している他家授精と自家授精の値の15個の差 $z_1 = 49$, $z_2 = -67, \ldots, z_{15} = -48$ と考えることにすれば，θ に対する推定量 $\theta(F_n)$ は $\bar{z}$ となる．図1(a)は，2000組のブートストラップ標本 $(z_1^*, \ldots, z_n^*)$ から計算された標本平均 $\hat{\theta}^* = \bar{z}^*$ の値のヒストグラムである．このとき，パーセンタイル法によるブートストラップ信頼区間を求めると，(0.73,



(a)



(b)

図 1．(a)2000組のブートストラップ標本から計算された $\bar{z}^*$ のヒストグラム．
(b)2000組のブートストラップ標本から計算された $\hat{r}^*$ のヒストグラム．

表3．ダーウィンのデータにおける，他家授精，自家授精の値の平均の差および平均の比に対する信頼係数95%のブートストラップ信頼区間．

	パーセンタイル法	abc 法	bc_a 法	t 法
$\mu_x - \mu_y$	(0.73, 38.20)	(-1.61, 36.36)	(-2.47, 36.53)	(-5.65, 40.84)
μ_x/μ_y	(1.01, 1.28)	(0.99, 1.27)	(0.99, 1.27)	(0.97, 1.31)

38.20）となる．これを，各z_iが$N(\theta,\ \sigma^2)$からの$i.i.d$標本であることを仮定する従来からの正規理論に基づく信頼区間，(2.48, 39.39）と比較すると，やや保守的（conservative）であることが分かる．さらに，パーセンタイル法を改善したabc法やbc_a法などによる信頼区間は，パーセンタイル信頼区間より区間幅が長く，またいずれも原点（帰無仮説と対応）を含んでいる．特に，ブートストラップ-t信頼区間は極めて長い(これは経験的によく知られていることであるが，証明はされていない)．また，ノンパラメトリック検定に基づく$\theta = \mu_x - \mu_y$に対する信頼区間については，竹内・大橋（(1981)，第4.5節)を参照されたい．

自家授精に対する他家授精の優越性の有無を検証するには，帰無仮説$H_0: r = 1$に対して対立仮説$H_1: r > 1$を検定する方法も考えられる．ただし，$r = \mu_x/\mu_y$である．これに対応して母平均の比 rに対する95%ブートストラップ信頼区間を作ってみよう．この場合初期標本を $z_1 =$ (188, 139)，$z_2 =$ (96, 163),..., $z_{15} =$ (96, 144）と考えれば，rの推定量は$\hat{r} = \bar{x}/\bar{y}$となる．

図1(b)は，2000組のブートストラップ標本 $(z_1^*, ..., z_n^*)$ から計算された$\hat{r}^* = \bar{x}^*/\bar{y}^*$の値のヒストグラムである．表3にまとめられている$r$に関する4種類の信頼区間は，母平均の差の場合と同様，ブートストラップ-t法が一番保守的であるのに対して，パーセンタイル法は他家授精の優越性を最も強く支持している．表3から分かるように，平均の差 $\hat{\theta} = \bar{x} - \bar{y}$ と平均の比 $\hat{r} = \bar{x}/\bar{y}$ によるブートストラップ解析では，ほぼ同様な結果が得られている．

これら4つの方法の中ではパーセンタイル法が，帰無仮説$H_0: \mu_x - \mu_y = 0$あるいは$\mu_x/\mu_y = 1$を一番棄却し易い．またブートストラップ-t法は区間幅が長く，一番保守的である．表中のブートストラップ信頼区間は2000組のブートストラップ標本に基づくものであり（abc法の場合は理論的に計算できる)，パーセンタイル法を除きEfron and Tibshirani ((1993), Appendix）にあるS言語で書かれた関数（電子メールで入手可能）を用いて計算されたものである．

上でみてきたように，ブートストラップ法による解析は極めて容易である．従来の正規分布表またはt分布表に相当する“ブートストラップ・パーセント点”の表は，計算機が自動的に作成してくれる．もちろんブートストラップ法の場合には，観測値が与えられたという条件のもとで計算が行われるので，計算機の生成する“ブートストラップ分布表”は，観測値に依存して決まるのである．

4. 仮説検定とブートストラップ法

4.1 はじめに

仮説検定の問題は，統計的誤差の推定や信頼区間の構成の問題とは，本質的に異なる点がある．仮説検定の問題では，データが得られた母集団の真の分布がどのようなものであるにせよ，我々が考察の対象とするのは，設定した帰無仮説のもとでの分布（以下帰無分布と呼ぶ）である．ブートストラップ法では，実際に観測されたデータに基づいて，条件付きで推論を行う．しかしその観測データは，厳密に考えれば，決して帰無仮説で想定した母集団分布から得られたものではな

いのである．そしてこのことが，仮説検定の問題にブートストラップ法を適用する際のネックになる．それにも拘らず，ブートストラップ法は仮説検定を行うための有力な方法であることを，本節で議論する．我々は以下で，2つの平均の同等性の検定を行うために，混合ブートストラップ検定という方法を提案する．ブートストラップ検定の一般的な議論については，Hinkley (1988, 1989) を参照のこと．また Romano (1988) は，距離に基づく検定のブートストラップ版について検討している (Beran and Ducharme (1991) も参照)．

以下の節では，まず従来からの検定とブートストラップ検定との違いを概観する．次に単純なブートストラップ法ではなぜうまく検定できないのか，その理由について議論する．そして混合ブートストラップ検定法を提案し，t 分布の近似という例を用いて，その信頼性を検証する．またこれ以外にも，いくつかのブートストラップ検定の方法を紹介する．最後にダーウィンのとうもろこしのデータに対して，混合ブートストラップ検定の手法を適用する．

4.2 これまでの検定とブートストラップ検定

仮説検定とは，例えば第1節の (1.1) で与えたような検定統計量の帰無分布に基づいて，標本空間を分割する手順のことである．従来から用いられてきた検定手順においては，標本は帰無仮説で設定した分布からランダムに取られていることを仮定して，議論が進められる．そして実際に観測されたデータに基づいて計算される検定統計量の実現値によって，帰無仮説が採択または棄却される．

これに対してブートストラップ検定では，帰無分布を近似するためにブートストラップ分布を作成し，この分布のパーセント点により標本空間の分割を行おうとする．ところでブートストラップ分布は，リサンプルから計算される検定統計量のヒストグラムによって置き換えられるため，

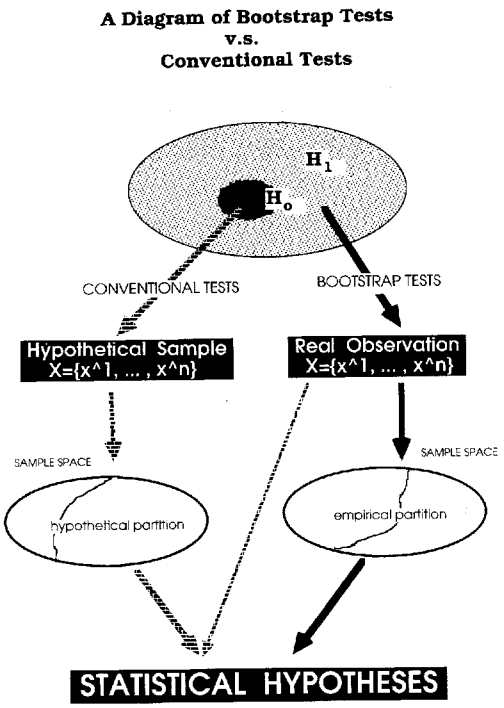


図2．ブートストラップ検定とこれまでの検定の比較．

異なる初期標本に対して，ブートストラップ近似分布（ヒストグラム）は異なるものとなり，したがって異なる標本空間の分割が得られる（詳しい議論については，以下の節を参照のこと）．図2は，これまでの検定とブートストラップ検定の違いを図式化して，比較を行ったものである．

4.3 単純なブートストラップ法適用の破綻

2標本問題に関する検定を考えてみよう．従来は議論を簡単にするために，分布に対する正規性の仮定と，等分散性の仮定をおくのが普通である．すなわち次が成り立っているような状況を考える．

$$X_1,\ldots,X_m \overset{iid}{\sim} N(\mu_x,\sigma^2),\qquad Y_1,\ldots,Y_n \overset{iid}{\sim} N(\mu_y,\sigma^2). \tag{4.1}$$

検定統計量を（1.1）とすれば，この場合 T の帰無分布は自由度 $m+n-2$ の t 分布となり，この問題は正確に解くことができる．

一方ブートストラップ検定では，観測データから得られるブートストラップ近似分布を利用して帰無分布を近似する．ところで統計的誤差の推定や信頼区間の構成の場合に述べたブートストラップ法の手順を考えれば（1.1）で定義された T の帰無分布を，ブートストラップ統計量 T^* の分布によって近似しようとするのが，もっとも単純な考え方であろう．なぜなら T と T^* との違いは，これらの値が計算される標本が，真の母集団分布から抽出されたものであるか，初期標本に基づく経験分布から抽出されたものであるかという点だけである．そして経験分布が真の母集団分布に対する合理的な近似であることを考えれば，T の分布を T^* の分布で近似するのは妥当であると思えるかもしれない．しかし我々の目的は，ブートストラップ法によって T の“帰無分布”を近似することである．一般に母集団の真の分布は帰無仮説のもとで想定する母集団分布とは異なっているから，それから抽出された初期標本に基づいて得られるブートストラップ近似分布は，T の帰無分布の合理的な近似とは考えられないのである．

このようにブートストラップ検定では，T の帰無分布は得られた初期標本に基づいて構成される．これに対して従来からの検定では，T の帰無分布を求めるために初期標本が使われることはない．初期標本は，検定統計量 T の実現値を計算するためにのみ用いられ，その値によって帰無仮説の採択，棄却が決定されるのである．この事情は図2にも示されているが，ブートストラップ検定と従来からの検定の重要な違いは，この「T の帰無分布を初期標本に基づいて構成するか否か」という点にある．そしてこの事がブートストラップ検定が汎用的なアルゴリズムとして記述可能であるという利点であると同時に，如何にして帰無分布を構成したらよいかという困難な問題が生じる原因ともなっている．次節で与える図3は，T の帰無分布と T^* のブートストラップ近似分布が大きくずれていることを示している例であり，従来からの検定とは異なり，どのようにして T の帰無分布を求めるのかが重大問題である．この点に関する詳しい議論は以下で行う．

4.4 混合ブートストラップ検定

上で述べたように，単純なブートストラップ検定が失敗する原因は，実際のデータが取られた2つの母集団の真の分布 $F(x)$，$G(x)$ が，帰無仮説のもとで対象としている分布の組の集合に入っていないからである．帰無仮説では，例えば $F(x)$ と $G(y)$ が同じであると仮定しても，得られたデータから作られた経験分布 $F_m(x)$ と $G_n(y)$ が近いとの根拠は何もない．いま，帰無仮説が $H_0:\mu_x=\mu_y$ で与えられる2標本検定問題を考えてみよう．2つの母集団分散が近似的に等しいと考えられる場合には次のようにすればよい．

1．まず2つの標本 $\boldsymbol{x},\boldsymbol{y}$ を混合する．すなわち

(4.2) $$\boldsymbol{z} = \{z_1,\ldots, z_{m+n}\} = \{x_1,\ldots,x_m, y_1,\ldots,y_n\}.$$

2．$\boldsymbol{z}$ から大きさ m, n の無作為標本 $\boldsymbol{x}_b^*$, $\boldsymbol{y}_b^*$ を復元抽出する．

$$\boldsymbol{x}_b^* = \{x_1^*,\ldots,x_m^*\} \Leftarrow \boldsymbol{z}$$

$$\boldsymbol{y}_b^* = \{y_1^*,\ldots,y_n^*\} \Leftarrow \boldsymbol{z}$$

3．$\boldsymbol{x}_b^*$, $\boldsymbol{y}_b^*$ を用いてブートストラップ統計量，T_b^* を計算する．

4．ステップ2と3を B 回繰り返し，$T_1^*,\ldots,T_B^*$ を計算する．

5．ステップ4から $\{T_b^*\}_{b=1}^B$ のヒストグラムを作成し，それを T の帰無分布のブートストラップ近似とする．

6．上で求めた T の帰無（近似）分布と，与えられる検定の有意水準とから，検定の棄却域を構成し，初期標本に基づいて計算された検定統計量 T の値が棄却域に入るか否かによって，帰無仮説 H_0 の棄却，採択を決定する．

上の検定手順を混合ブートストラップ検定と呼ぶことにする．混合ブートストラップ検定の妥当性を検証するために，(4.1) のような正規性，等分散性の仮定が成立しており，$m = n = 5$ であるような状況を考えよう．このとき検定統計量 (1.1) の帰無分布は自由度8の t 分布であり，我々はこの分布をなるべく精度よく近似したい訳である．さて実際に初期標本が抽出される2つの母集団分布が，それぞれ $N(1, 1)$ および $N(0, 0.9^2)$ であったとしよう．これら2つの母集団から，それぞれ大きさ5の初期標本 $x_1,\ldots,x_5$ および $y_1,\ldots,y_5$ を抽出し，混合ブートストラップ検定法を適用してみる．ブートストラップ・リサンプリングの回数は200回とする．

この計算結果をまとめたものが図3および表4で，上記のような数値実験を100回繰り返して行った場合の平均値である．図3は，混合ブートストラップ検定法，前節でのべた単純なブートストラップ検定法および次節で述べる位置調整ブートストラップ検定法による近似分布を，帰無分布（自由度8の t 分布）と比較した図である．また表4は，種々のブートストラップ検定法によって帰無分布を近似した場合の，分布の裾の部分における相対誤差の値を示している．この表から，2つの初期標本を混合するという素朴なアイディアが非常に有効に機能していることが分かる．

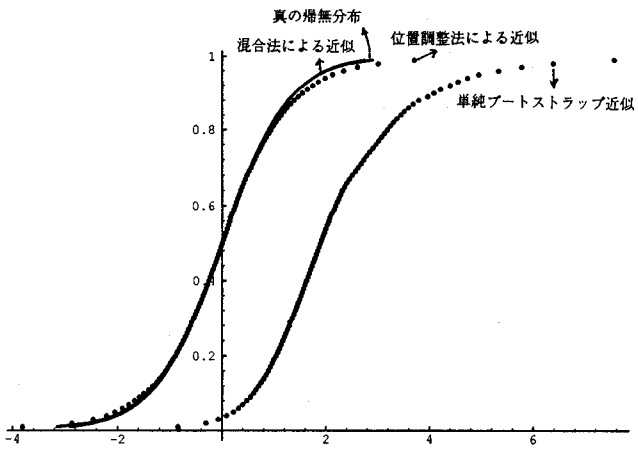


図3．3種類の検定方法におけるブートストラップ近似分布の比較．

4.5 その他のブートストラップ検定

我々は前節で，等分散性の仮定のもとで，2つの母平均が等しいか否かを検定するための，混合ブートストラップ検定法を提案した．2つの標本を混合するという考え方は，帰無仮説の構造をデータに反映させる1つのうまいやり方であろう．実際 Boos et al.（1989）では，ここで述べた標本の混合という考え方が既に用いられており，そのアイディアに基づいて尺度の均一性に対するプールド・ブートストラップ検定法が議論されている．帰無仮説の構造をデータに反映させるための調整は，他にもいくつか提案されている．母平均の検定の場合については，Efron and Tibshirani ((1993), p. 224) に，次のような方法が示唆されている．すなわち初期標本を $x_1,\ldots, x_m$ および $y_1,\ldots,y_n$ とするとき，これらを $x_i' = x_i - \bar{x} + \bar{z}$ および $y_i' = y_i - \bar{y} + \bar{z}$ と変換し，この $\{x_1',\ldots,x_m'\}$ および $\{y_1',\ldots,y_n'\}$ からブートストラップ標本を抽出するという方法である．ただし $\bar{x}, \bar{y}$ はそれぞれの標本の平均であり，$\bar{z}$ はプールされた標本の平均である．4.4節で述べた数値実験において単純なブートストラップ検定法の場合にこの方法を適用した結果が，表4の第5行目に与えられている．この数値を見ると，この方法によっては調整が十分にうまく行われているとは言い難いことが分かる．次に考えられるのは，位置と尺度の両方を考慮する変換であろう．すなわち上の x_i, y_i を $x_i' = (x_i - \bar{x})/s_x$, $y_i' = (y_i - \bar{y})/s_y$ によって変換する方法である．ただし s_x, s_y は，それぞれの標本の標準偏差である．単純なブートストラップ検定法に対するこの場合の計算結果は，表4の第6行目に与えられており，位置だけの変換の場合より，特に上側の裾の部分で改善が著しい．表4の第2，3行目には，混合ブートストラップ法に対して位置の変換を行った場合，および位置と尺度の変換を行った場合の結果が与えられている．ただし位置／尺度の変換は，標本を混合する前に行っている．これらを表の第1行目の変換を行わなかった場合と比較してみると，位置／尺度変換による改善度はほとんどないか，ごく僅かであることが分かる．

4.6 ダーウィンのデータに対するブートストラップ検定

表1のダーウィンのとうもろこしのデータに関して，3節と同様に15組の対を作って，自家授精に対する他家授精の優越性の有無を検定してみよう．帰無仮説を $H_0: \mu_x = \mu_y$，対立仮説を $H_1: \mu_x > \mu_y$ とし，前節までで述べた各種のブートストラップ検定法により，片側検定を行う．

まず標本の混合を行わないブートストラップ検定で，位置変換を行う方法を適用したところ，片側達成有意水準（検定統計量の実現値に対する上側確率；one-sided achieved significance

表4．6種類のブートストラップ検定法による帰無分布の裾の部分に対する近似．

検定法	パーセント点	1%	2%	3%	4%	5%	95%	96%	97%	98%	99%
混合	変換なし	.09	.04	.03	.03	.03	.01	.01	.02	.03	.05
	位置混合	.09	.02	.03	.03	.03	.02	.02	.01	.01	.03
	位置 - 尺度混合	.10	.04	.03	.03	.03	.00	.00	.00	.01	.01
単純	変換なし	.71	.87	.97	1.04	1.12	1.66	1.66	1.64	1.61	1.61
	位置	.31	.17	.12	.10	.10	.14	.17	.19	.23	.28
	位置 - 尺度	.25	.15	.10	.09	.08	.03	.05	.07	.10	.15

註：表中の値は，|（真値－近似値）／真値| で定義される相対誤差の値を表わしている．上の3段は混合ブートストラップ検定，下の3段は単純なブートストラップ検定の場合に対応している．また上記2種類の各場合の第1，2，3行目は，それぞれデータ変換を行わなかった場合，位置変換のみを行った場合，位置および尺度変換を行った場合に対応している．

level）は0.043であった．これに対して位置変換を行わない混合ブートストラップ検定法の場合の達成有意水準は0.012であった．これより，混合ブートストラップ検定法を適用した場合の方が，「他家授精と自家授精の間には差がない」という帰無仮説を棄却し易いことが分かる．

次に標本の混合を行うブートストラップ検定では，位置変換を行った場合の達成有意水準は0.006にまで減少し，2種類のとうもろこし間の差が小さな有意水準でも検出できていることになる．位置-尺度変換を行った場合の達成有意水準は0.011であり，位置だけの変換の場合より検出力は低下している．しかし対応する標本の混合を行わない場合の達成有意水準0.015よりは小さくなっている．以上の達成有意水準の値は，いずれの場合も2000組のブートストラップ標本に基づいて計算されたものである．比較のために，従来から用いられてきたいくつかのノンパラメトリック検定の結果を与えておくと，メディアン検定，ウィルコックソン検定，並べかえ検定の場合の両側達成有意水準の値は，それぞれ0.001, 0.003, 0.024である（竹内・大橋（1981）参照）．この問題に関しては，母集団の真の構造がわからないため，各種の検定法の有効性についての議論はできないが，達成有意水準の値でみる限り，あまり大きな差は観察されなかった．

5. おわりに —— リサンプリング推定方程式の観点から

本特集「計算統計学の発展」の趣旨を考慮して，本論文ではブートストラップ法に関する多くの重要な話題を割愛してきた．そこで終わりの本節では，これまでに述べなかったブートストラップ法の他の重要な発展について，主として推定方程式の観点から簡単に触れてみることにしよう．

推定方程式に基づく推定の考え方（Godambe (1960, 1976, 1985)）は，統計的推定論において最もよく使われてきた2つの手法，すなわちガウスの最小自乗法とフィッシャーによる最尤法の考え方を融合させたものとみることができよう．これに関しては，生物統計学，確率過程論，標本調査論などの分野で，いくつかの発展があった（Godambe (1991) 参照）．

我々は第2節，第3節において，統計的誤差の推定量（2.2），(2.3）や信頼区間（3.1）に関するブートストラップの手法を，分かり易く実際的な方法で紹介してきた．しかし推定方程式の観点からブートストラップ法を理解しておくと，より深い考察が可能となるかもしれない．いま，次式における $f_t(F_0, F_1)$ を適当に定め，興味をもつパラメータが次の t に関する方程式の解として定義できたとしよう．

$$E_{F_0} f_t(F_0, F_1) = 0 \tag{5.1}$$

このとき，上で定義したパラメータに対するブートストラップ推定値は，次の標本方程式（Hall (1992), p. 6）と呼ばれる方程式の解になる．

$$E_{F_1} f_t(F_1, F_2) = 0 \tag{5.2}$$

ここで F_0 は未知の母集団分布，F_1 は初期標本に基づく経験分布（前節まででは F_n と書いたもの），F_2 はブートストラップ標本に基づく経験分布である．また f_t は，添字 t に依存する汎関数である．もし $f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) - t$ とすれば，標本方程式の解として（2.2）のブートストラップ・バイアス推定量が得られる．興味のあるパラメータが（5.1）の解になるような $f_t(F_0, F_1)$ の関数型を見つけるのが困難な場合もあるが，前節までの統計的誤差や分布関数のような期待値の形で表せる問題はすべてこのような定式化が可能である．

次に，次式で与えられる線形推定方程式（Godambe (1985)）と呼ばれる方程式に対するブートストラップ法の適用について考えてみよう

(5.3)
$$\sum_{i=1}^{n} g_i(\boldsymbol{X}, \theta) = 0.$$

ここで$\boldsymbol{X} = (X_1,...,X_n)$は確率変数ベクトルで，各要素が互いに独立である必要はない．またθは興味のある実数値パラメータである．この例として，マルコフ過程に対する条件付き最小自乗推定方程式が考えられるが，その場合には$g_i(\boldsymbol{X}, \theta)$は次式で与えられる．

$$g_i(\boldsymbol{X}, \theta) = \{X_i - E(X_i|X_{i-1}, \theta)\}\{\partial E(X_i|X_{i-1}, \theta)/\partial\theta\}.$$

いま（5.3）の解が$\hat{\theta}$であるとしよう．この場合，ジャックナイフ法による$\hat{\theta}$のバイアスおよび分散の推定量は，それぞれ（2.2），（2.3）で与えられる（Lele (1991a)）．ただしj番目の観測値を除いた場合の推定値$\hat{\theta}_{(j)}$は，次式を解くことによって求めなければならない．

$$\sum_{i \neq j} g_i(\boldsymbol{X}, \theta) = 0$$

次にθに対するブートストラップ信頼区間について考えると，これも容易に構成することができる．いま$y_i = g_i(\boldsymbol{X}, \hat{\theta})t_i$, $(i = 1,...,n)$と定義する．ここで各t_iは，平均0，分散1をもつ*i.i.d.*確率変数である．このとき，θに対する信頼係数$1-2\alpha$の信頼区間は，次式によって与えられる．

$$[\hat{\theta} + w(\alpha)/I(\hat{\theta}),\ \hat{\theta} + w(1-\alpha)/I(\hat{\theta})]$$

ただし$I(\hat{\theta}) = |\sum_i \partial g_i(\boldsymbol{X}, \hat{\theta})/\partial\theta|/n$である．また$w(\beta)$は，初期標本が$(y_1,...,y_n)$であるときの標本平均$\bar{y}$のブートストラップ分布の$100\cdot\beta$パーセント点である．ここでポイントとなるのは$\{t_i\}$の列の適切な選び方であるが，この点に関してはLele (1991b) を参照のこと．

このように推定方程式を用いることによって，*i.i.d.*でない場合の統計的推測を行う際にも，リサンプリング法が有用な手段として適用できる可能性が広がりつつある．実際，現在では*i.i.d.*の場合におけるブートストラップ法のアルゴリズムをさらに工夫するよりは，*i.i.d.*でない場合のブートストラップ法の適用に関する研究が盛んになりつつある（Young (1994) の最近の展望論文，特に第5節を参照）．

ブートストラップ法の研究は，現在でもなお急速に発展を続けている．理論的に興味のある研究としては，エッジワース展開 (Hall (1992), Booth and Hall (1994) 等) やノンパラメトリック尤度 (Owen (1988, 1990), Hinkley (1988), Efron and Tibshirani (1993), Chapter 24) に関するものが挙げられよう．またその適用分野も広がっており，生物統計学への適用 (Moulton and Zeger (1989, 1991), Carr and Portier (1993) and Hjorth (1994))，モデル選択のようなより複雑な問題への適用 (LePage and Billard (1992))，判別分析 (Efron (1983, 1986), 小西・本多 (1992))，因子分析 (Ichikawa and Konishi (1995))，情報量規準への適用 (Konishi and Kitagawa (1995)) などがある．

謝　辞

本論文の作成に際し，レフェリーの方に貴重なコメントを頂きました．特に論文の読み易さ，参考文献の引用のしかたについて改善が計られたと思います．深く御礼申し上げます．

参　考　文　献

Beran, R. and Ducharme, G. R. (1991). *Asymptotic Theory for Bootstrap Methods in Statistics*, Les

Publications Centre de Recherches Mathematiques, Université de Montréal, Montréal.
Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap, *Ann. Statist.*, **9**, 1196-1217.
Boos, D., Janssen, P. and Veraverbeke, N. (1989). Resampling from centered data in the two-sample problem, *J. Statist. Plann. Inference*, **21**, 327-345.
Booth, J. and Hall, P. (1994). On the validity of Edgeworth and saddlepoint approximations, *J. Multivariate Anal.*, **51**, 121-138.
Carr, G. J. and Portier, C. J. (1993). An evaluation of some methods for fitting dose-response models to quantal-response developmental toxicology data, *Biometrics*, **49**, 779-791.
Darwin, C. (1876). *The Effects of Cross- and Self-fertilisation in the Vegetable Kingdom*, John Murray, London.
DiCiccio, T. T. and Romano, J. P. (1995). On bootstrap procedures for second-order accurate confidence limits in parametric models, *Statistica Sinica*, **5**, 141-160.
Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Statist.*, **7**, 1-26.
Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.
Efron, B. (1983). Estimating the error rate of a prediction rule: improvements on cross-validation, *J. Amer. Statist. Assoc.*, **78**, 316-331.
Efron, B. (1986). How biased is the apparent error rate of a prediction rule?, *J. Amer. Statist. Assoc.*, **81**, 461-470.
Efron, B. (1987). Better bootstrap confidence intervals (with discussion), *J. Amer. Statist. Assoc.*, **82**, 171-200.
Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, London.
Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer, New York.
Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, *Philos. Trans. Roy. Soc. London Ser. A*, **222**, 309-368.
Fisher, R. A. (1960). *The Design of Experiments*, 7th ed., Oliver and Boyd, Edinburgh.
Freedman, D. A. (1981). Bootstrapping regression models, *Ann. Statist.*, **9**, 1218-1228.
Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation, *Ann. Math. Statist.*, **31**, 1208-1211.
Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations, *Biometrika*, **63**, 277-284.
Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes, *Biometrika*, **72**, 419-428.
Godambe, V. P. (ed.)(1991). *Estimating Functions*, Clarendon Press, Oxford.
Hall, P. (1992). The bootstrap and Edgeworth expansions, *Springer Ser. Statist.*, Springer, New York.
Hinkley, D. V. (1988). Bootstrap methods (with discussion), *J. Roy. Statist. Soc. Ser. B*, **50**, 321-337.
Hinkley, D. V. (1989). Bootstrap significance tests, *Proceedings of 47th Session of the International Statistical Institute, Paris, 29 August-6 September 1989*, Vol. 3, 65-74.
Hjorth, J. S. U. (1994). *Computer Intensive Statistical Methods*, Chapman and Hall, London.
Ichikawa, M. and Konishi, S. (1995). Application of the bootstrap methods in factor analysis, *Psychometrika*, **60**, 77-93.
Johnson, N. and Kotz, S. (1970). *Continuous Univariate Distributions*, Vol. 2, Houghton Mifflin, Boston.
小西貞則・本多正幸(1992). 判別分析における誤判別率推定とブートストラップ法, 応用統計学, **21**, 67-100.
Konishi, S. and Kitagawa, G. (1995). Generalized information criterion and the bootstrap, Research Memo., No. 549, The Institute of Statistical Mathematics, Tokyo.
Lele, S. (1991a). Jackknifing linear estimating equations: asymptotic theory and applications in stochastic processes, *J. Roy. Statist. Soc. Ser. B*, **53**, 253-267.
Lele, S. (1991b). Resampling using estimating equations. *Estimating Functions* (ed. V. P. Godambe), 295-304, Clarendon Press, Oxford.
LePage, R. and Billard, L. (ed.)(1992). *Exploring the Limits of Bootstrap*, Wiley, New York.
松縄 規 (1994). 分布の起源——ノンパラメトリックな統計的不確定性関係と統計基礎方程式, 統計数理, **42**, 197-214.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.
Moulton, L. H. and Zeger, S. L. (1989). Analyzing repeated measures on generalized linear models via the bootstrap, *Biometrics*, **45**, 381-394.
Moulton, L. H. and Zeger, S. L. (1991). Bootstrapping generalized linear models, *Comput. Statist. Data Anal.*, **11**, 53-63.
Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, **75**, 237-249.
Owen, A. B. (1990). Empirical likelihood confidence regions, *Ann. Statist.*, **18**, 90-120.
Quenouille, M. H. (1956). Notes on bias in estimation, *Biometrika*, **43**, 353-360.
Reid, N. (1994). A conversation with Sir David Cox, *Statist. Sci.*, **9**, 439-455.
Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests, *J. Amer. Statist. Assoc.*, **83**, 698-708.
Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, **9**, 1187-1195.
竹内　啓, 大橋靖雄 (1981).『統計的推測 —— 2標本問題』, 日本評論社, 東京.
Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika*, **61**, 439-447.
Young, G. A. (1994). Bootstrap: more than a stab in the dark?, *Statist. Sci.*, **9**, 382-415.

Bootstrap Method
—— An Introduction from a Two Sample Problem

Jin Fang Wang
(The Institute of Statistical Mathematics)
Masaaki Taguri
(Department of Mathematics and Informatics, Chiba University)

Bootstrap is a data-based simulation method proposed by Efron in 1979. It is a method, based on observed data sets, for assessing statistical properties such as bias, variance and prediction error. More complicated and well understood applications include construction of confidence limits, estimation of distribution functions, etc. The power and versatility of the bootstrap lie mainly in the fact that the bootstrap resampling process is automatically achieved through computers.

This paper attempts to give an easy and practical introduction to the bootstrap method through a historical two sample problem of Darwin. The bootstrap, through various examples, is shown to be an easy and valid method for varieties of statistical problems, which deserves to be added into every practical statistician's toolbox.

Some new ideas on bootstrap tests are also discussed. We propose the mixed bootstrap tests for testing the equality of two means when the variances are nearly homogeneous. The purpose of introducing bootstrap tests is twofold: the bootstrap is both useful and less naive in other more delicate problems. Current state of developments are also briefly reviewed, with resampling estimating equations in particular.

Key words: Bias, bootstrap, confidence interval, estimating equation, jackknife, mixed bootstrap test, nonparametric maximum likelihood, two-sample problem, variance.