

Numerical examination on bootstrap tests for difference between two distributions using Kolmogorov-Smirnov statistic

北海道大学・大学院工学研究科 桜井 裕仁 (Hirohito Sakurai)
Graduate School of Engineering, Hokkaido University

筑波大学・数学系 高橋 邦彦 (Kunihiko Takahashi)
Institute of Mathematics, University of Tsukuba

概要

本論文では、ブートストラップ法による2つの母集団分布の有意差検定法を提案する。検定統計量には Kolmogorov-Smirnov 統計量を用い、その帰無分布の近似法として、5種類のリサンプリング法を議論する。さらに2群のデータの標本数が極めてアンバランスな場合を想定し、それらの検定のサイズ、検出力の検討を数値的に行う。

1 はじめに

連続な分布 $F(x)$, $G(x)$ から互いに独立に抽出された2組の標本 $\mathbf{X} = \{X_1, \dots, X_n\}$, $\mathbf{Y} = \{Y_1, \dots, Y_m\}$ に基づいて、2つの分布が等しいという仮説

$$H_0 : F(x) = G(x) \quad \text{for all } x \tag{1}$$

を検定する問題は、2標本問題と呼ばれる。ただし、 $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F$, $Y_1, \dots, Y_m \stackrel{i.i.d.}{\sim} G$ とする。(1)の問題を扱うノンパラメトリック検定の中でも、Kolmogorov-Smirnov 統計量

$$T(\mathbf{X}, \mathbf{Y}) = \sup_{-\infty < x < \infty} |\hat{F}_n(x) - \hat{G}_m(x)| \tag{2}$$

を用いる Kolmogorov-Smirnov 検定 (以下では、K-S 検定と略す) は有名である。ここで、 $\hat{F}_n(x)$, $\hat{G}_m(x)$ は経験分布関数を表し、

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\}, \quad \hat{G}_m(x) = \frac{1}{m} \sum_{j=1}^m I\{Y_j \leq x\},$$

である。ただし、 $I\{\cdot\}$ は定義関数である。

このような検定問題は、次のような現実の問題に対して利用できるであろう。例えば、ある薬の効果が既にいくつかの国や地域の多くのデータによって確かめられ使用されている場合に、別な国においても、その薬の効果が認められるか否かを確かめたいとする。ここで一般に、調査費用をはじめとする種々の制約により、十分な大きさの標本数を得ることは困難であるので、もし既に得られている大量のデータの情報を利用できるのであれば、いかに少ないデータによってその薬効を確かめることができるかということが問題になる。このような場合には、既に得られている大量のデータの分布 F と、新たに得られた少ないデータの分布 G が等しいか否かという検定問題、すなわち、

$$\text{帰無仮説 } H_0 : F = G, \quad \text{対立仮説 } H_1 : F \neq G \tag{3}$$

を考えればよいであろう (Akahira and Takahashi, 2001). ところがこれまでに, K-S 検定に関して, 2 群の標本数が極端に異なる場合 ($n \gg m$) の挙動についてはあまり調べられていない. さらに, 検定統計量 (2) を用いる場合の検出力などについても, より一般的な状況のもとで解析的に導出するのは困難である.

そこで本論文では, $n \gg m$ の場合の K-S 検定の数値的な挙動を調べることを目的とする. また, 検定統計量 (2) の帰無仮説のもとでの分布 (以下では, 帰無分布と呼ぶ) の近似法として, ブートストラップ法を用いる場合の K-S 検定法を提案する. ブートストラップ法によれば, 統計数値表 (例えば, 山内編 (1972) を参照) にない場合をも扱うことができ, さらには, 解析の手順を自動化できる利点がある.

検定問題 (3) をブートストラップ法の観点から扱っている文献には, Præstgaard (1995) や Jing (2000) などがある. Præstgaard (1995) では, 1 次元データの場合の (3) の問題を, ブートストラップ検定と並べ替え検定により検討している. 一方, Jing (2000) では, Præstgaard (1995) のブートストラップ検定を多次元データの場合に拡張し, さらに, 標本が $k (\geq 2)$ 個の場合の検定問題も扱っている.

ところで, ブートストラップ法を仮説検定の問題に適用する場合には, Hall and Wilson (1991) で指摘されているように, 「帰無仮説を反映するようリサンプリング」を行わなければならない. それは, 以下の理由のためである. 我々の目的は, ブートストラップ法によって検定統計量 (2) の帰無分布を近似することである. しかし一般には, 母集団の真の分布は, 帰無仮説で想定する母集団分布とは異なっているので, 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, (2) の帰無分布の合理的な近似とは考えられない (汪・田栗, 1996). したがって, 本論文で扱う検定問題 (3) の場合にも, 帰無仮説の構造をうまく反映するようリサンプリングを行わなければならない.

ここで, ブートストラップ法による 2 群の平均値の有意差検定を行う場合に, 帰無分布の近似に用いられる代表的な方法には,

- (A) 平均調整をした個々のサンプルからリサンプリングする方法 (Efron and Tibshirani, 1993)
- (B) 2 標本を混合し, 混合した標本からリサンプルを抽出する方法 (汪・田栗, 1996)

がある. 上述した Præstgaard (1995) や Jing (2000) では, いずれも (B) の方法によりリサンプルを発生させている.

そこで本論文では, まず, 上記の (A) に対応する検定法を提案する. 次に, 汪・田栗 (1996) と同様な考え方により, 上記の (B) の方法に対応する, Præstgaard (1995) の方法を若干修正する方法を提案する.

本論文の構成は以下の通りである. 2.1 節では, まず, 上記 (A) に対応する検定法を提案する. 次に 2.2 節では, (B) に対応する Præstgaard (1995) の検定方法を概観し, その修正法を提案する. 3 節では, 2 節の検定法による検定のサイズ, 検出力を, 2 群の標本数がアンバランスな場合をシミュレーションにより検討する. 最後に, 4 節では, 3 節の結果から得られる知見, および今後の課題をまとめる.

2 検定のアルゴリズム

本節では、大別して2種類のブートストラップ検定法を提案する。2.1節、および2.2節では、それぞれ1節で述べた(A), (B)に対応する検定法を述べる。以下で提案する手法を分類すると、表1のようになる。

表 1: 検定手法の分類

	調整しない	位置調整	位置・尺度調整
(A)	×	Algorithm 2.1*	Algorithm 2.2*
(B)	Algorithm 2.3 (Præstgaard, 1995)	Algorithm 2.4*	Algorithm 2.5*

*印が本節で提案する検定法

2.1 位置・尺度調整ブートストラップ検定

Efron and Tibshirani (1993)では、2群の母平均のブートストラップ有意差検定を行う場合に、帰無仮説を反映するリサンプリング法として、初期標本の平均を調整した後に各標本からリサンプルを発生させる方法を議論している。本節ではこの方法を、2群の分布が等しいか否かという検定問題(3)の場合に適用することを提案する。検定の手順を Algorithm 2.1にまとめる。

Algorithm 2.1 (位置調整ブートストラップ検定)

1. 有意水準 α を設定する。
2. 初期標本 $\mathbf{x} = \{x_1, \dots, x_n\}$, $\mathbf{y} = \{y_1, \dots, y_m\}$ に基づき、検定統計量の実現値 $t_{obs} = T(\mathbf{x}, \mathbf{y})$ を計算する。
3. $\tilde{\mathbf{x}} = \{x_1 - \bar{x}, \dots, x_n - \bar{x}\}$, および $\tilde{\mathbf{y}} = \{y_1 - \bar{y}, \dots, y_m - \bar{y}\}$ を計算する。ただし、 $\bar{x} = \sum_{i=1}^n x_i/n$, $\bar{y} = \sum_{j=1}^m y_j/m$ である。
4. $\tilde{\mathbf{x}}$ から大きさ n のリサンプル $\mathbf{x}^{*b} = \{x_1^{*b}, \dots, x_n^{*b}\}$ を、 $\tilde{\mathbf{y}}$ から大きさ m のリサンプル $\mathbf{y}^{*b} = \{y_1^{*b}, \dots, y_m^{*b}\}$ を無作為復元抽出し、 $t^{*b} = T(\mathbf{x}^{*b}, \mathbf{y}^{*b})$ ($b = 1, \dots, B$) を計算する。
5. 手順4を B 回繰り返す、

$$\begin{cases} H_0 : \text{reject,} & \text{if } \widehat{\text{ASL}}_{boot} \leq \alpha, \\ H_0 : \text{accept,} & \text{if } \widehat{\text{ASL}}_{boot} > \alpha \end{cases} \quad (4)$$

により、帰無仮説の棄却、採択を決定する。ただし、 $\widehat{\text{ASL}}_{boot} = \sum_{b=1}^B I\{t^{*b} \geq t_{obs}\}/B$ は達成有意水準 (achieved significance level) である。

次に, Algorithm 2.1 に対する修正法を提案する. Algorithm 2.1 では, 帰無仮説の構造をデータに反映させるために, 2 群のデータの標本平均が一致するような調整を行った. しかし, 2 群のデータの尺度が異なる場合には, 両者の尺度が等しくなるような調整も行っておくべきであろう. Algorithm 2.2 では, 位置・尺度の両方を調整する検定法を示す.

Algorithm 2.2 (位置・尺度調整ブートストラップ検定)

Algorithm 2.1 の手順 1, 2, 5 は同様にし, 手順 3, 4 を以下のように変更する.

3° $\tilde{\mathbf{x}}' = \{(x_1 - \bar{x})/s_x, \dots, (x_n - \bar{x})/s_x\}$, および $\tilde{\mathbf{y}}' = \{(y_1 - \bar{y})/s_y, \dots, (y_m - \bar{y})/s_y\}$ を計算する. ただし, $s_x = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)}$, $s_y = \sqrt{\sum_{j=1}^m (y_j - \bar{y})^2 / (m-1)}$ である.

4° $\tilde{\mathbf{x}}'$ から大きさ n のリサンプル $\mathbf{x}^{*b} = \{x_1^{*b}, \dots, x_n^{*b}\}$ を, $\tilde{\mathbf{y}}'$ から大きさ m のリサンプル $\mathbf{y}^{*b} = \{y_1^{*b}, \dots, y_m^{*b}\}$ を無作為復元抽出し, $t^{*b} = T(\mathbf{x}^{*b}, \mathbf{y}^{*b})$ ($b = 1, \dots, B$) を計算する.

2.2 位置・尺度調整混合ブートストラップ検定

本節では, 1 節で述べた (B) に対応する方法, すなわち, 2 つの初期標本を混合し, 混合した標本からのリサンプリングを行う方法を議論する. まず, Præstgaard (1995) による検定法を Algorithm 2.3 にまとめる.

Algorithm 2.3 (混合ブートストラップ検定 (Præstgaard, 1995))

1. 有意水準 α を設定する.
2. 初期標本 $\mathbf{x} = \{x_1, \dots, x_n\}$, $\mathbf{y} = \{y_1, \dots, y_m\}$ に基づき, 検定統計量の実現値 $t_{obs} = T(\mathbf{x}, \mathbf{y})$ を計算する.
3. 2 つの初期標本 \mathbf{x}, \mathbf{y} を混合し, $\tilde{\mathbf{w}} = \{x_1, \dots, x_n, y_1, \dots, y_m\}$ とおく.
4. $\tilde{\mathbf{w}}$ から大きさがそれぞれ n と m のリサンプル $\mathbf{x}^{*b} = \{x_1^{*b}, \dots, x_n^{*b}\}$, $\mathbf{y}^{*b} = \{y_1^{*b}, \dots, y_m^{*b}\}$ を無作為復元抽出し, $t^{*b} = T(\mathbf{x}^{*b}, \mathbf{y}^{*b})$ ($b = 1, \dots, B$) を計算する.
5. 手順 4 を B 回繰り返す, (4) と同様にして, 帰無仮説の棄却, 採択を決定する.

次に, Algorithm 2.3 に対する修正法を提案する. 汪・田栗 (1996) で述べられているように, 単純な混合ブートストラップ検定の場合よりも, 2 つの標本を混合する前に, データの位置調整, または位置と尺度の両方の調整を行うことにより, 僅かにではあるが, 検定統計量の裾の部分の近似を改善できる可能性がある.

そこでここでは, データの位置調整, および位置と尺度の両方の調整を行う検定法を提案する. 前者を Algorithm 2.4 に, 後者を Algorithm 2.5 にまとめる.

Algorithm 2.4 (位置調整混合ブートストラップ検定)

Algorithm 2.3 の手順 1, 2, 5 は同様にし, 手順 3, 4 を以下のように変更する.

- 3'. $\tilde{\mathbf{x}}' = \{x_1 - \bar{x}, \dots, x_n - \bar{x}\}$, および $\tilde{\mathbf{y}}' = \{y_1 - \bar{y}, \dots, y_m - \bar{y}\}$ を計算し, $\tilde{\mathbf{w}}' = \{x_1 - \bar{x}, \dots, x_n - \bar{x}, y_1 - \bar{y}, \dots, y_m - \bar{y}\}$ とおく.
- 4'. $\tilde{\mathbf{w}}'$ から大きさがそれぞれ n と m のリサンプル $\mathbf{x}^{*b} = \{x_1^{*b}, \dots, x_n^{*b}\}$, $\mathbf{y}^{*b} = \{y_1^{*b}, \dots, y_m^{*b}\}$ を無作為復元抽出し, $t^{*b} = T(\mathbf{x}^{*b}, \mathbf{y}^{*b})$ ($b = 1, \dots, B$) を計算する.

Algorithm 2.5 (位置・尺度調整混合ブートストラップ検定)

Algorithm 2.3 の手順 1, 2, 5 は同様にし, 手順 3, 4 を以下のように変更する.

- 3''. $\tilde{\mathbf{x}}'' = \{(x_1 - \bar{x})/s_x, \dots, (x_n - \bar{x})/s_x\}$, および $\tilde{\mathbf{y}}'' = \{(y_1 - \bar{y})/s_y, \dots, (y_m - \bar{y})/s_y\}$ を計算し, $\tilde{\mathbf{w}}'' = \{(x_1 - \bar{x})/s_x, \dots, (x_n - \bar{x})/s_x, (y_1 - \bar{y})/s_y, \dots, (y_m - \bar{y})/s_y\}$ とおく.
- 4''. $\tilde{\mathbf{w}}''$ から大きさがそれぞれ n と m のリサンプル $\mathbf{x}^{*b} = \{x_1^{*b}, \dots, x_n^{*b}\}$, $\mathbf{y}^{*b} = \{y_1^{*b}, \dots, y_m^{*b}\}$ を無作為復元抽出し, $t^{*b} = T(\mathbf{x}^{*b}, \mathbf{y}^{*b})$ ($b = 1, \dots, B$) を計算する.

3 検定のサイズ, 検出力の数値的検討

本節では, 2 群の母集団分布 F, G が, それぞれ正規分布 $N(0, 1), N(\mu, \sigma^2)$ の場合のブートストラップ検定 Algorithm 2.1~2.5 による検定のサイズ, 検出力をシミュレーションにより検討する. ここで, 有意水準 α は, $\alpha = 0.05, 0.10$ の 2 通りの場合を考え, 各母集団分布から大きさがそれぞれ n と m ($n \gg m$) の標本を独立に 1000 組発生させる. ただし, 2 群の標本数 (n, m) , および分布のパラメータ (μ, σ^2) は, 以下のすべての組み合わせを考え, 各場合において, 検定統計量 (2) の帰無分布の近似を行うために, $B = 2000$ 回のリサンプリングを行う.

- $n = 1000, m = 20, 50$ (2 通り)
- (μ, σ^2) の組み合わせ
 - Model (a): $\sigma^2 = 1, \mu = 0.0 (0.1) 1.0$ (11 通り)
 - Model (b): $\sigma^2 = 2, \mu = 0.0, 0.2, 0.5$ (3 通り)
 - Model (c): $\sigma^2 = 4, \mu = 0.0, 0.2, 0.5$ (3 通り)

以下ではまず, 各検定法の検定のサイズに関する検討を行う. 表 2 は, Model (a) の場合の検定のサイズをまとめたものである. リサンプリング法が同じである表 2 の 1 行目と 2 行目 (Algorithm 2.1, 2.2) 間の比較, および 3~5 行目 (Algorithm 2.3, 2.4, 2.5) 間の比較を行うと, 大きな差は見られなかった. すなわち, この場合には, 初期標本に対して行った位置調整や位置・尺度調整の効果は見られなかった. しかし, 帰無分布の近似法の違いによって, 次のような特徴的な傾向が観察された. 各々の標本からリサンプリングする検定法 (Algorithm 2.1, 2.2) では, 名目上の水準を過小評価するのに対し, 2 群のデータを混合してからリサンプリングする検定法 (Algorithm 2.3, 2.4, 2.5) では, 名目上の水準を維持する傾向が見られた.

次に、各検定法の検出力に関する検討を、2群のデータの分散が等しい場合とそうでない場合とに分けて行う。まず、図1, 2は、それぞれ有意水準 α が5%, 10%で2群のデータの分散が等しい場合、すなわち、Model (a)に対する検出力の結果をまとめたグラフである。ただし、各図の縦軸と横軸は、それぞれ検出力と μ の値である。図1 (i), (ii), および図2 (i), (ii)に対応する検出力の詳細な値は、それぞれ表3, 4, および表5, 6に与える。この場合には、Algorithm 2.1~2.5に対応する5つの検出力曲線が2つのグループに分かれている。各図において、検出力が高くなっている曲線は、Algorithm 2.3~2.5によって得られた検出力曲線であり、それよりも下にあるのが、Algorithm 2.1, 2.2によって得られた検出力曲線である。ここで表7~表10の1, 2行目、および3~5行目に着目すると、検定のサイズの場合と同様に、各グループに対応する検定法による検出力はほぼ同じである。ゆえに、この場合にも、初期標本に対して行った位置調整や位置・尺度調整の効果は観察されなかった。

一方、2群のデータの分散が異なる場合、すなわち、Model (b), (c)に対する検出力の結果をグラフにまとめると、図3~図6のようになる。図3, 4, 5, 6に対応する検出力の詳細な値は、それぞれ表7, 8, 9, 10に与える。この場合には、Algorithm 2.1~2.5に対応する5つの検出力曲線が3つのグループに分かれるという結果が得られた。各図において、検出力が1番高くなっている曲線は、Algorithm 2.3~2.5によって得られたものである。ここで表7~表10の3~5行目に着目すると、この3つの検定法による検出力はほぼ同じである。したがって、Algorithm 2.3~2.5によって得られた検出力曲線はほぼ重なり、1つの曲線のように見える。その次に検出力の高い曲線は、Algorithm 2.2によって得られたものであり、検出力が1番低い曲線は、Algorithm 2.1によって得られたものである。

2群のデータの分散が異なる場合、各々の標本からリサンプリングを行う検定法では、初期標本に対して尺度の調整を行うことの有用性が顕著に表れている。例えば、図5 (ii)の場合、初期標本に対して位置調整のみを行うAlgorithm 2.1による検出力は、 $\mu = 0, 0.2, 0.5$ では0.018, 0.039, 0.198であるが、位置・尺度調整を行うAlgorithm 2.2による検出力は、それぞれ0.545, 0.640, 0.839となる。よって、各々の標本からリサンプリングを行う検定法の場合には、リサンプリングする前に、位置調整だけでなく、位置・尺度の両方の調整を初期標本に対して行うべきであろう。

表 2: Empirical level of bootstrap tests in simulation study for Model (a)

	$\alpha = 0.05$		$\alpha = 0.10$	
	m		m	
	20	50	20	50
Algorithm 2.1	0.014	0.011	0.034	0.028
Algorithm 2.2	0.014	0.011	0.032	0.031
Algorithm 2.3	0.071	0.060	0.126	0.115
Algorithm 2.4	0.070	0.060	0.125	0.114
Algorithm 2.5	0.072	0.060	0.126	0.115

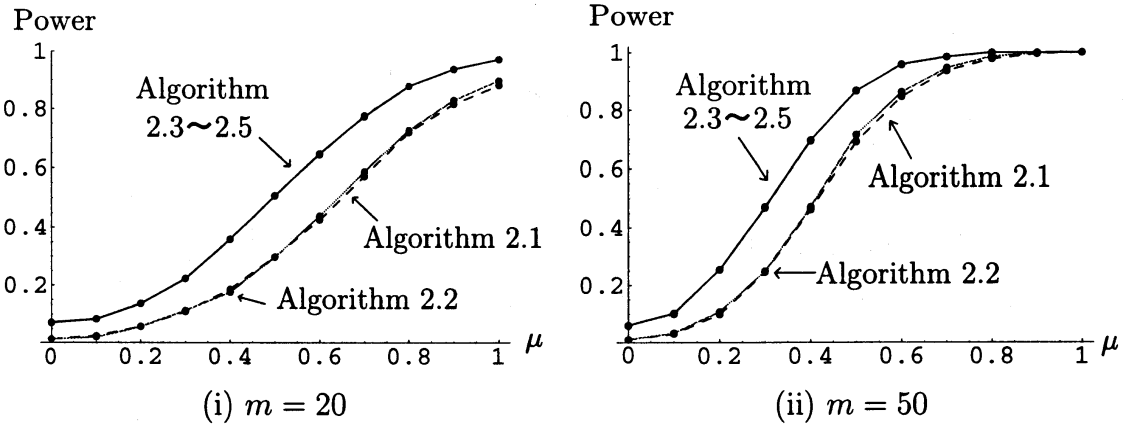


图 1: Comparison of empirical powers of 0.05 level bootstrap tests for Model (a)

表 3: Empirical power of 0.05 level bootstrap tests for Model (a) with $m = 20$

	μ									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Algorithm 2.1	0.026	0.057	0.109	0.186	0.298	0.436	0.572	0.719	0.816	0.881
Algorithm 2.2	0.021	0.057	0.113	0.176	0.297	0.439	0.588	0.726	0.830	0.898
Algorithm 2.3	0.084	0.138	0.222	0.358	0.508	0.644	0.777	0.878	0.937	0.972
Algorithm 2.4	0.082	0.137	0.222	0.359	0.507	0.647	0.773	0.878	0.941	0.971
Algorithm 2.5	0.083	0.136	0.223	0.359	0.507	0.647	0.773	0.879	0.938	0.971

表 4: Empirical power of 0.05 level bootstrap tests for Model (a) with $m = 50$

	μ									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Algorithm 2.1	0.031	0.098	0.245	0.462	0.689	0.845	0.935	0.976	0.994	0.999
Algorithm 2.2	0.034	0.108	0.248	0.471	0.713	0.861	0.945	0.985	0.996	0.999
Algorithm 2.3	0.100	0.253	0.468	0.693	0.867	0.956	0.984	0.998	0.999	1.000
Algorithm 2.4	0.103	0.254	0.471	0.695	0.865	0.958	0.983	0.999	0.999	1.000
Algorithm 2.5	0.103	0.254	0.471	0.696	0.866	0.958	0.984	0.999	0.999	1.000

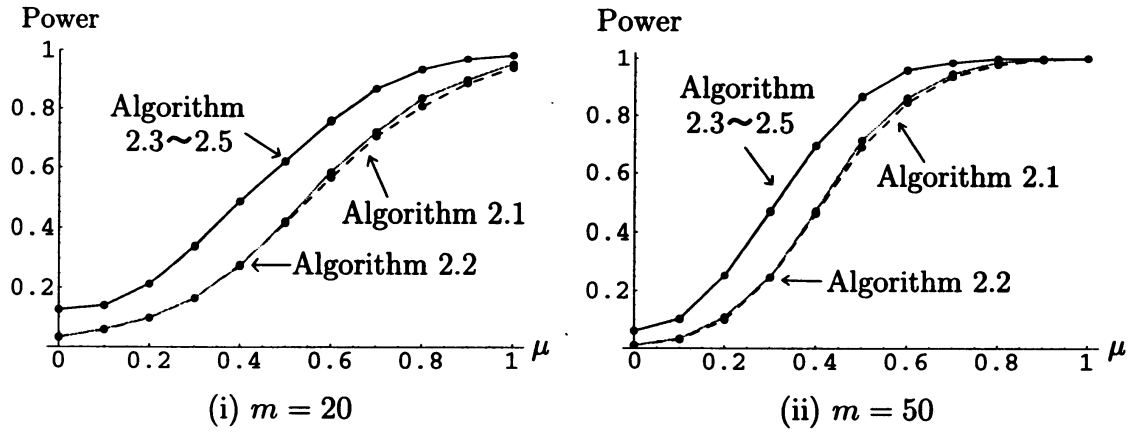


图 2: Comparison of empirical powers of 0.10 level bootstrap tests for Model (a)

表 5: Empirical power of 0.10 level bootstrap tests for Model (a) with $m = 20$

	μ									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Algorithm 2.1	0.057	0.098	0.164	0.279	0.415	0.564	0.705	0.807	0.885	0.938
Algorithm 2.2	0.060	0.099	0.166	0.275	0.421	0.584	0.720	0.835	0.898	0.951
Algorithm 2.3	0.139	0.212	0.338	0.487	0.619	0.759	0.867	0.932	0.969	0.982
Algorithm 2.4	0.141	0.214	0.341	0.487	0.620	0.756	0.867	0.932	0.969	0.982
Algorithm 2.5	0.140	0.214	0.342	0.487	0.620	0.756	0.866	0.932	0.969	0.982

表 6: Empirical power of 0.10 level bootstrap tests for Model (a) with $m = 50$

	μ									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Algorithm 2.1	0.060	0.158	0.372	0.596	0.797	0.911	0.964	0.992	0.999	1.000
Algorithm 2.2	0.067	0.177	0.378	0.616	0.815	0.920	0.975	0.995	0.999	1.000
Algorithm 2.3	0.197	0.356	0.590	0.793	0.918	0.976	0.994	0.999	1.000	1.000
Algorithm 2.4	0.200	0.358	0.591	0.791	0.919	0.975	0.995	0.999	1.000	1.000
Algorithm 2.5	0.201	0.358	0.589	0.794	0.918	0.975	0.995	0.999	1.000	1.000

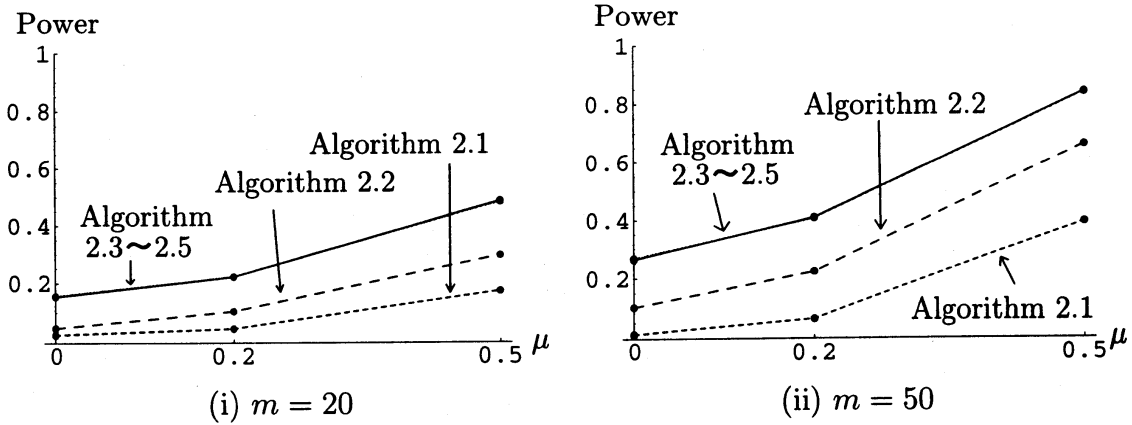


图 3: Comparison of empirical powers of 0.05 level bootstrap tests for Model (b)

表 7: Empirical power of 0.05 level bootstrap tests for Model (b)

	$m = 20$			$m = 50$		
	$\mu = 0.0$	$\mu = 0.2$	$\mu = 0.5$	$\mu = 0.0$	$\mu = 0.2$	$\mu = 0.5$
Algorithm 2.1	0.026	0.045	0.172	0.013	0.063	0.393
Algorithm 2.2	0.047	0.103	0.295	0.102	0.222	0.660
Algorithm 2.3	0.156	0.222	0.482	0.269	0.407	0.838
Algorithm 2.4	0.155	0.221	0.484	0.266	0.407	0.840
Algorithm 2.5	0.153	0.221	0.486	0.264	0.409	0.840

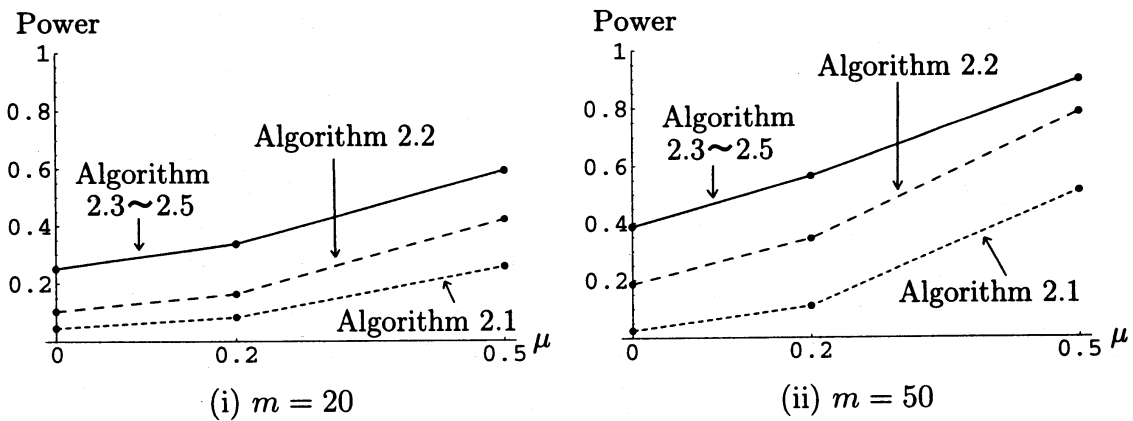


图 4: Comparison of empirical powers of 0.10 level bootstrap tests for Model (b)

表 8: Empirical power of 0.10 level bootstrap tests for Model (b)

	$m = 20$			$m = 50$		
	μ			μ		
	0.0	0.2	0.5	0.0	0.2	0.5
Algorithm 2.1	0.046	0.083	0.256	0.025	0.109	0.511
Algorithm 2.2	0.104	0.164	0.421	0.186	0.344	0.784
Algorithm 2.3	0.251	0.338	0.591	0.387	0.565	0.895
Algorithm 2.4	0.251	0.336	0.593	0.386	0.564	0.897
Algorithm 2.5	0.251	0.337	0.593	0.388	0.564	0.896

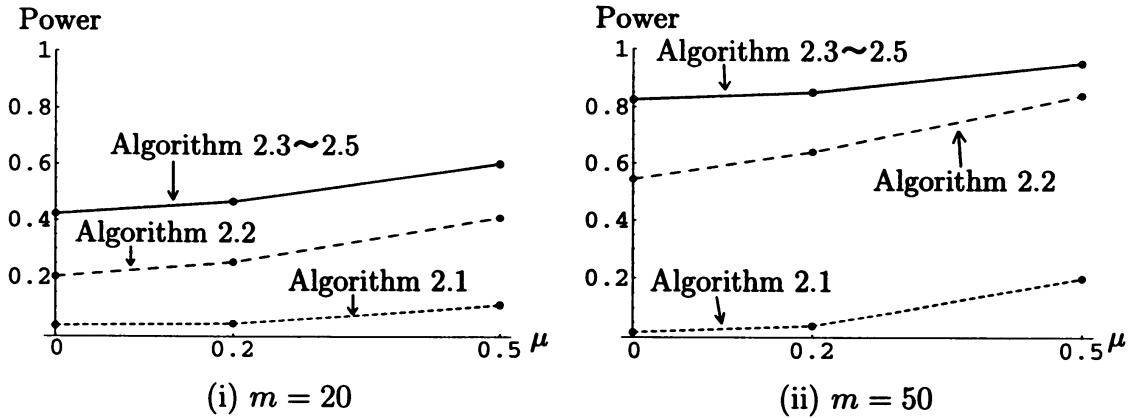


图 5: Comparison of empirical powers of 0.05 level bootstrap tests for Model (c)

表 9: Empirical power of 0.05 level bootstrap tests for Model (c)

	$m = 20$			$m = 50$		
	μ			μ		
	0.0	0.2	0.5	0.0	0.2	0.5
Algorithm 2.1	0.037	0.043	0.108	0.018	0.039	0.198
Algorithm 2.2	0.203	0.253	0.410	0.545	0.640	0.839
Algorithm 2.3	0.425	0.468	0.600	0.825	0.852	0.950
Algorithm 2.4	0.424	0.467	0.602	0.824	0.850	0.949
Algorithm 2.5	0.423	0.466	0.602	0.827	0.849	0.950

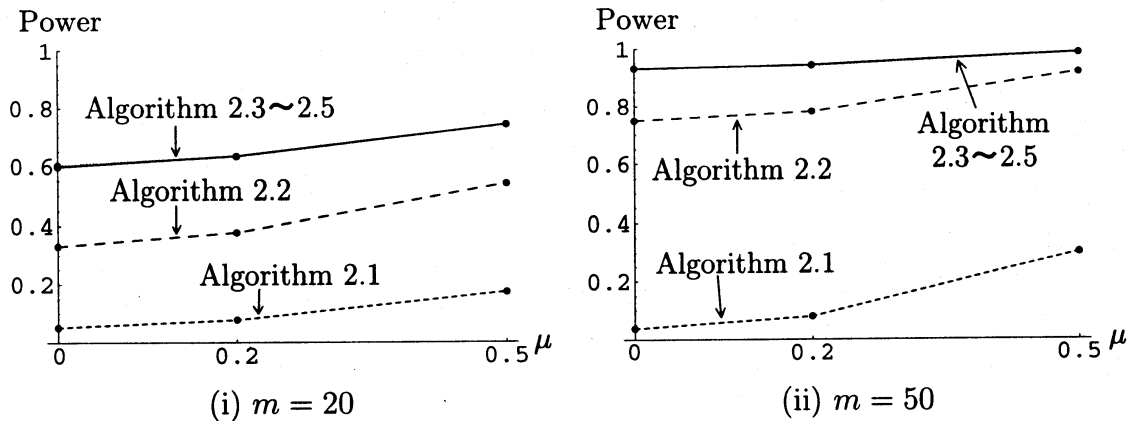


図 6: Comparison of empirical powers of 0.10 level bootstrap tests for Model (c)

表 10: Empirical power of 0.10 level bootstrap tests for Model (c)

	$m = 20$			$m = 50$		
	μ	μ	μ	μ	μ	μ
	0.0	0.2	0.5	0.0	0.2	0.5
Algorithm 2.1	0.052	0.077	0.173	0.037	0.079	0.301
Algorithm 2.2	0.332	0.377	0.539	0.750	0.781	0.917
Algorithm 2.3	0.603	0.633	0.741	0.932	0.943	0.984
Algorithm 2.4	0.600	0.633	0.742	0.932	0.942	0.984
Algorithm 2.5	0.598	0.634	0.742	0.931	0.944	0.985

4 まとめと今後の課題

本論文では、検定統計量には Kolmogorov-Smirnov 統計量 (2) を用い、その帰無分布の近似法として 5 種類のブートストラップ法を適用した場合の 2 つの母集団分布の有意差検定法を議論した。3 節では、2 群のデータの標本数が極めてアンバランスな場合を想定し、2 群の母集団分布が正規分布の場合の検定のサイズ、検出力の検討を数値的に行った。

全体的な傾向として、各群のデータからリサンプリングする方法よりも、2 群のデータを混合した標本からリサンプリングする方法の方が、名目上の検定のサイズを維持し、かつ検出力が高くなるという結果が得られている。したがって、もし 2 群のデータが今回検討したような分布に従う場合には、Algorithm 2.3~2.5 を用いるべきであろう。しかし、若干検出力は低くなるものの、第 1 種の過誤を抑えたいという場合には、Algorithm 2.2 の適用も考えられよう。また、実際のデータ解析に適用する場合には、2 群のデータの尺度均一性が保証されない可能性があり、そのような場合には、Algorithm 2.5、または Algorithm 2.2 を用いばよいであろう。

3節の数値的検討は、2群の母集団分布が非常に限られた状況でしか行われていない。今後は、例えば2群の母集団分布が、標準正規分布と形状の似ている自由度が3または4の t 分布、あるいはベータ分布の場合に、どの程度の差を検出することができるのか検討の必要があると考える。また、Præstgaard (1995)では並べ替え検定法も提案されており、この手法と今回提案した手法間の比較・検討も必要であろう。

ところで今回は、初期標本を混合する方法では、位置調整、位置・尺度調整の有用性は確認できなかった。この点についての検討は今後の課題であると考え。さらに、従来のK-S検定とリサンプリング法を適用した場合の検定との検定のサイズ、検出力の比較を行い、2群の平均値の有意差検定の場合に用いられている1節の(A)、(B)の方法が、本研究の対象とする検定問題(3)の場合にも適用可能か否かを検討することも今後の検討課題であると考え。

以上では、2群の標本数が極めてアンバランスな場合の検討を行ったが、最後に、2群の標本数が同じ場合の検定のサイズの検定結果を、表11にまとめる。この場合には、3節と同様に、Algorithm 2.2, 2.4, 2.5では、初期標本に対して行った位置調整や位置・尺度調整の効果は見られなかった。また、各々の標本からリサンプリングする検定法(Algorithm 2.1, 2.2)では、名目上の水準を過小評価するのに対し、2群のデータを混合してからリサンプリングする検定法(Algorithm 2.3, 2.4, 2.5)では、名目上の水準を維持する傾向が見られる。Præstgaard (1995)やJing (2000)では、2群の標本数が等しい場合についても、その検定のサイズ、検出力に関する数値的挙動は明らかにされていない。したがって、2群の標本数が同じ場合にも、上述したような2群の標本数がアンバランスな場合と同様な検討が必要であろう。

表 11: Empirical level bootstrap tests for Model (a) with $n = m$

	$\alpha = 0.05$ ($n = m$)					$\alpha = 0.10$ ($n = m$)				
	20	30	50	100	200	20	30	50	100	200
Algorithm 2.1	0.012	0.006	0.010	0.006	0.005	0.024	0.019	0.024	0.014	0.011
Algorithm 2.2	0.013	0.009	0.012	0.006	0.006	0.029	0.026	0.025	0.016	0.015
Algorithm 2.3	0.039	0.055	0.049	0.049	0.043	0.104	0.092	0.114	0.093	0.076
Algorithm 2.4	0.042	0.055	0.051	0.048	0.041	0.102	0.091	0.116	0.093	0.076
Algorithm 2.5	0.041	0.052	0.051	0.048	0.041	0.102	0.092	0.116	0.093	0.075

参考文献

- [1] Akahira, M. and Takahashi, K. (2001). A simulation study on testing the hypothesis in the two sample problem, 数理解析研究所講究録, 1224, 182–186.
- [2] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

- [3] Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing, *Biometrics*, **47**, 757–762.
- [4] Jing, P. (2000). Bootstrap tests for the equality of distributions, *Korean Journal of Computational and Applied Mathematics*, **7**, 347–362.
- [5] Præstgaard, J. T. (1995). Permutation and bootstrap Kolmogorov-Smirnov tests for the equality of two distributions, *Scandinavian Journal of Statistics*, **22**, 305–322.
- [6] 汪 金芳, 田栗 正章 (1996). ブートストラップ法 — 2 標本問題からの考察, *統計数理*, **44**, 3–18.
- [7] 山内 二郎 編 (1972). 統計数値表 JSA-1972, 日本規格協会.