外国語教育メディア学会 (LET) 関西支部 メソドロジー研究部会 2014 年度 第 5 号報告論集
草薙 邦広 (pp. 1–15)

外国語教育研究における
ブートストラップ法の応用可能性

草薙 邦広
名古屋大学大学院生
日本学術振興会特別研究員

概要

本稿の目的は，外国語教育研究におけるブートストラップ法の応用可能性について述べることである。ブートストラップ法は，頑健統計の一部をなす手法のひとつであり，主にさまざまな母数の区間推定に用いられる。本稿では，まず頑健統計に関する一般的な概説，ブートストラップ法やそれに類似する手法の原理についての紹介をおこなう。その後，外国語教育研究データの分析を視野にいれながら，ブートストラップ法を用いた分析方法の例をいくつか示す。

Keywords: 統計手法，ブートストラップ法，信頼区間，研究方法，頑健統計

1. 背景

現在，国内の外国語教育研究の大半が量的手法を用いている（e.g., Mizumoto, Urano, & Maeda, 2014）。しかしながら，その量的研究手法の洗練性は望ましいものとはいいがたい。たとえば，統計的仮説検定においては，検定力が十分でないものも実際に数多く見られる（草薙・水本・竹内, 2014）。そしてそれは，標本サイズの決定手順がえてして十分でないからとも考えられる（草薙他, 2014）。また，基礎的な統計手法の選択や，統計量の報告にも問題があるという指摘もかなり前からなされている（e.g., 前田, 2000）。

しかし，このような問題は国内に限ったことではない。当該分野の国際的トップジャーナル掲載論文においても同様の指摘が数多くなされている（e.g., Plonsky, 2013, 2014; Plonsky & Gass, 2011）。同時に，外国語教育研究において研究方法論（メソドロジー）に関する関心が高まっているともいわれている（e.g., Plonsky, 2014）。特に，国内外を問わず心理学などを代表として，さまざまな分野において頑健統計（robust statistics）の手法を援用する機運が高まっている（全体的な概観としては，大久保・岡田, 2012; 応用言語学においては，Larson-Hall, 2012; Larson-Hall & Herrington, 2010 など）。

頑健統計がなにかということを理解するためには，まず，頑健性（robustness）という性質について理解するべきである。頑健性とは，一般的にいって「ある統計手法が仮定

している条件を満たしていなくとも結果が妥当である程度」を示す。頑健統計は，このような性質をもつ統計手法全般であり，正規分布に強く依存する従来の統計手法に対して，さまざまな確率分布を代用したり，分布に依存せずに妥当な結果を出すような手法を示す場合が多い。

たとえば，平均値（mean. *M*）という統計量は，外れ値の影響を受けやすいということが広く知られている（e.g., 南風原, 2002）。これをもって代表値とみなすことが不適切である場合には，刈り込み平均（trimmed mean）や中央値（median）を代用するとよい。また，標準偏差（standard deviation, *SD*）に対しては四分位区間（IQR）を代用することもある。正規性が満たされないデータに対する統計的仮説検定においては，従来からノンパラメトリック検定と呼ばれる手法が代用される。これらの手法は，「頑健性をもつ」といってさしつかえない。

近年では，統計的仮説検定に依存しすぎない研究のあり方も模索されてきている。たとえば，国内でも効果量，検定力，信頼区間の使用について普及に努める試みも多く見られる（e.g., 水本・竹内, 2008, 2011）。また，研究成果の体系的統合ともいえるメタ分析の重要性，追行研究，シミュレーション研究の必要性も主張されている。加えて，研究に関わるデータの適切な可視化方法について，その技術的向上の必要性も認知されるところである（e.g., 草薙, 2014d）。

このような動きの背景には，研究ツールの発達もあるだろう。現在は，標準的なパーソナルコンピュータ上で複雑な統計分析をおこなうことができる無償のツールが開発され，広く流通している（R などがその最たるものである）。また，パーソナルコンピュータの機能の向上によって，計算資源が相対的に安価になってきている。さらに，統計手法の解説書や，インターネット上の資料など，統計手法に関する学習機会も充実してきていることが要因として考えられるだろう。

外国語教育研究においても，各種統計手法に関わる分析環境は向上しているが，依然として以下のような問題に苛まれる場合が多い：(a) しばしば教育現場を対象とするため，理想的な実験環境が得にくく，現実による制約が多い（e.g., 草薙, 2014c），(b) 分析上，統計手法におけるさまざまな仮定が満たせない場合も多い（e.g., 標本サイズ，正規性，過誤の制御）。このような背景において，頑健統計は今後の研究におけるひとつの重要な観点になるかもしれない。本稿では，特にその中でもブートストラップ法（bootstrapping）について焦点をあてて，その外国語教育における応用可能性について考察を加える。

2. ブートストラップ法の概観

ブートストラップ法とは，頑健統計の一角であり，再標本化法（resampling）のひとつとみなすことができる（Efron & Tibshirani, 1993; 汪，2003）。再標本化法とは，一度得

た標本から，分析上再度標本を作り直す手法一般を示す。ここでは，便宜上最初に得た標本を元標本，再度標本としてケースを選び出して作った標本をサブ標本（subsample）と呼ぶ。図 1 に，母集団，元標本およびサブ標本の関係を図示する。

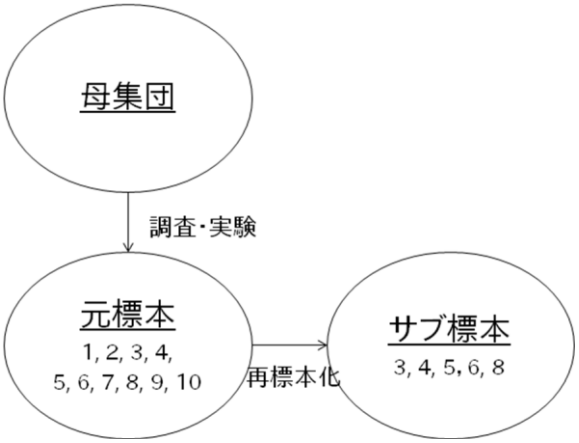

図 1. 母集団，元標本およびサブ標本の関係性

また，ブートストラップ法はモンテカルロ法と呼ばれる手法の性質ももっている。モンテカルロ法とは，一般にシミュレーションなどによって乱数を複数生成する過程を経る統計手法を指す。

再標本化法には，さまざまなタイプの手法がある。一つ目の種類は，「パラメトリックブートストラップ」と呼ばれるものである。この手法は，経験分布（元標本から推定された母数）から乱数を生成することによってサブ標本を得る。たとえば，実験で 10 人の標本（元標本）を得て，平均値および標準偏差をもとめたとき（仮に $M = 10, SD = 1$），そのような母数をもつ分布にもとづく乱数を 10 個作るとする。それらの乱数が，11, 9, 9, 11, 8, 10, 11, 10, 10, 11 といった値になったとしよう。このような乱数は，分布とその母数（元標本による推定）が正しい場合，母集団から新しく得たケースと同様となるはずである。この手法も広い意味で再標本化とみなす。

二つ目の方法は，「ノンパラメトリックブートストラップ」と呼ばれるものである。この手法では，経験分布から得られる乱数を用いるのではなく，元標本からケースを抽出する。図 1 の例では，元標本から 3, 4, 5, 6, 8 というケースを抽出している。この手法にはさらに，「復元抽出」と呼ばれる手法があり，それは元標本からケースをひとつ抽出したときに，再度そのケースを元標本から抽出することを許す（元に戻す，復元する）手続きをとる。つまり，原理上，図 1 の例では，3, 3, 3, 3, 3 といったサブ標本ができる可能性がある。本稿では，主にこの手法についてくわしく紹介する。

ここで紹介する最後の方法は，ジャックナイフ法（ジャックナイフ推定）と呼ばれるものである。ジャックナイフ法では，元標本のサイズを n としたとき，n -1 のサイズか

らなるサブ標本を *n* 組生成する。この手法は復元抽出なし，と捉えられる。本稿ではくわしく述べないが，この手法は相関係数における外れ値による影響の緩和などに応用される。この手法を用いた国内の外国語教育における研究論文には草薙（2014a）などがある。

ブートストラップ法は，主に母数の区間推定に用いられる。ここからは，例としてノンパラメトリックブートストラップを用いた母平均の区間推定（信頼区間）について紹介する。まず，仮に以下のような元標本を得たとする（表 1）。このデータをあるテストの得点と考える。

表 1
元標本の得点（*n* = 12）

ケース	得点
学生 1	69
学生 2	58
学生 3	67
学生 4	63
学生 5	87
学生 6	68
学生 7	65
学生 8	58
学生 9	88
学生 10	77
学生 11	75
学生 12	71

この標本の平均値は 70.50，標準偏差は 9.82 である。この標本から，ブートストラップ法を用いて母平均値の区間推定をする。まずは，サブ標本（この場合，特にブートストラップ標本とも呼ぶ）のサイズを決める。ここでは元標本と同じサイズである 12 とする。つまり，*n* = 12 となるブートストラップ標本を復元抽出を用いて元標本から生成する。次に，ブートストラップ標本の数を決める。ここでは仮に 1,000 とする。[1] ブートストラップ標本数は通常，*B* と表せられる。このようにブートストラップ標本のサイズと *B* を決定し，元標本のデータから 12 個ずつのケースを復元抽出し，1,000 組作り出すという手続きをする。

このようにブートストラップ標本を膨大な数作り出し，各ブートストラップ標本の任意の統計量を得る。たとえば，ここでは平均値とする。つまり，ブートストラップ標本 $_1$ の平均値は 71.21，ブートストラップ標本 $_2$ の平均値は 73.10 というように 1,000 個につ

いてすべて計算するのである。このブートストラップ標本ごとに計算した統計量をひとつの標本と見立てるとき，その分布は母数の確率分布に近似すると考えられる。図 2 にブートストラップ標本における平均値の分布をヒストグラムで示す。

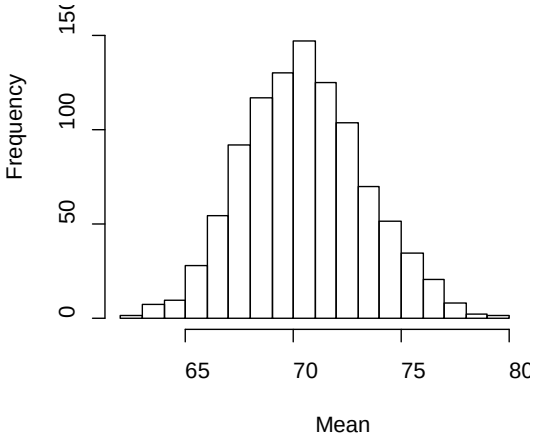

図 2. ブートストラップ標本における平均値の分布

この分布について，中央値を求めると，70.46 となる。これは，元標本のみによる平均値（母平均の推定値）よりも母数の真値に近いと考えられる。また，この分布についての 95%のパーセンタイル範囲は，母数の 95%信頼区間とみなすことができる。ちなみに，このデータでは[65.25, 76.34]となった。この区間推定の方法は，最も単純なもので，一般にパーセンタイル法と呼ばれる。

母平均の区間推定をおこなうだけであれば，通常の正規分布を用いた計算で十分である場合が多いが，ブートストラップ法を用いた区間推定は，分布を仮定せず，さまざまな統計量に対して応用できる。たとえば，*t* 検定の *p* 値，効果量，標準化偏回帰係数などである。本稿のおよぶ範囲ではないが，区間推定には，上記のパーセンタイル法のほかにも，ベーシック法，BCa 法，ブートストラップ *t* 法など，さまざまな方法がある（汪，2003 に詳しい）。ブートストラップ法はシミュレーションをともなう手法であるため，試行毎に値が微動することがある。この点には十分注意すべきである。

ブートストラップ法の利点として，その汎用性の高さと，特に確率分布といった前提への依存の少なさがあげられる。また，区間推定のみならず，点推定値も一般に元標本のみによる推定値よりも母数に近いと考えられるため，特に小標本の場合などにおいて，標本誤差を緩和する効果をもつ。

ブートストラップ法を用いた分析をおこなうことができるツールにはさまざまなものがある。たとえば，商用の製品では，SPSS，Amos，MATLAB などがあげられる。また，R のパッケージには *boot*，*simpleboot*，*bootstrap* といったパッケージ [2] が公開されている。Microsoft 社の Excel をプラットフォームとする無償の統計解析ツール HAD においてもブ

ートストラップ法を用いた分析手法が実装されている。また，ブートストラップ法を用いた分析は計算が莫大になるものの，アルゴリズム自体は非常に単純であるため，R のデフォルト関数を用いても容易に計算することができる（附録にスクリプトの例を示す）。

3. さまざまな応用例

ここからは，外国語教育研究を文脈としながらさまざまな応用例をあげていく。

3.1. 小標本における中央値の信頼区間

仮に 8 人の実験協力者にテストを実施したとする。図 3 にそのヒストグラムを示す。標本の分布を確認したところ，正規性が十分とはいえず，代表値として中央値を用いることにした。ここで，中央値の 95%信頼区間を求めたい。ブートストラップ法のほかにも中央値の信頼区間を求める方法はあるもの，ブートストラップ法は間便であるため，代用することとした。この（元）標本の中央値は 38.50 であった。

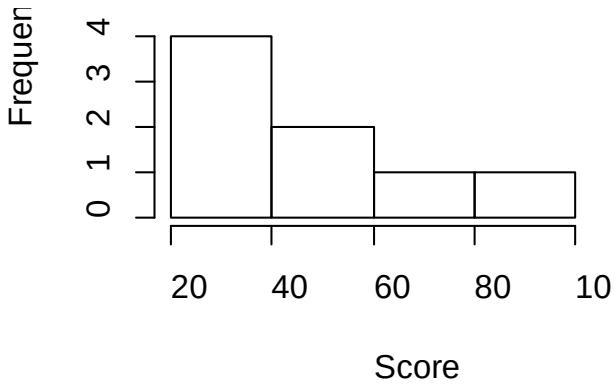

図 3. 3.1 における元標本のヒストグラム

パーセンタイル法を用いて，B = 1,000，ブートストラップ標本サイズを元標本と等しい 8 とした場合における中央値の 95%信頼区間を求めると，[21.98, 76.00]となった。また，点推定値は 38.70 であった。

3.2. *t*検定におけるブートストラップ法

ブートストラップ法を用いた検定には実にさまざまな種類がある（e.g., Larson-Hall & Herrington, 2010; Plonsky, Egbert, & LaFlair, 2014; Wilcox, 2003）。ここでは，例として，二標本対応なしの t 検定について紹介する。まず，以下のようなふたつの元標本をもつとする（表 2）。仮に男女ごとにおける英語のテストの得点としよう。この元標本のデータに対して，等分散性を仮定しない Welch の方法による t 検定をおこなうと，$t(37) = 2.059$，$p = .046$ となり，統計的に差は有意である。

表 2

元標本における記述統計

	標本サイズ	平均値	標準偏差	尖度	歪度
男性グループ	22	69.97	9.22	-0.35	-0.33
女性グループ	22	62.85	13.36	0.30	-1.03

注. このデータは乱数を用いた作成したものである。

Plonsky et al.（2014）と同様に，それぞれのグループから復元抽出をしてブートストラップ標本を生成する。B = 1,000，ブートストラップ標本サイズを元標本と等しい 22 とする。これら 1,000 ずつのブートストラップ標本の組に対して，1,000 回の t 検定をおこなう。このときの検定統計量 t を対象として，95%信頼区間を求め，その下限と固定値としての 0 を比較する方法を取る。t 検定では，帰無仮説において $t = 0$ と仮定されているからである。仮に今回のブートストラップ法による t 値の推定区間における下限が 0 を下回るならば，元標本のみによる検定結果は，第一種の過誤である可能性があると考えられる。なお，今回の区間推定にもパーセンタイル法を用いた。

ブートストラップ法による区間推定の結果，t 値の下限は 0.14，そして上限は 4.38 であった。ブートストラップ標本における t 値の分布を図 4 に示す。

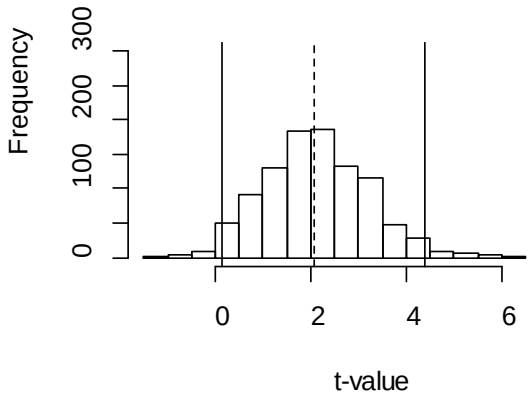

図 4. ブートストラップ標本における t 値の分布を示す。縦に走る実線は 95%信頼区間の上限と下限，破線は p = .05 となる点を示している。

この例では，t 値の下限が 0 を下回らないため，元標本による t 検定の結果が第一種の過誤を犯していると直ちにはいいがたい。しかしこれは他にも重要な情報を示している。まず，図 3 の分布からみて，5%水準で有意となる t 値，2.06 ほどが分布の山の頂点となっていることである。これでは，検定結果を十分な根拠をもって採用することができない。

次に，今回の（1,000 組のブートストラップ標本に対する）1,000 回の検定の p 値の

累積確率を図 5 に示す。ここから単純に計算すると, 1%水準において統計的に平均差が有意となったブートストラップ標本の数は, 294 組（全体の 29%ほど）であり, 5%水準では, 526 個（53%ほど）である。

このように, ブートストラップ法を援用した検定手法は, 研究で得たデータを従来の方法よりもより客観的に評価することができると考えられる（Plonsky et al., 2014）。

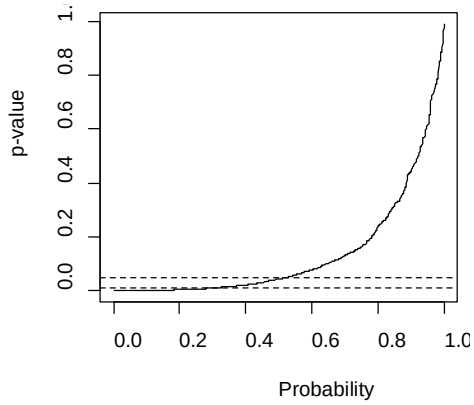

図 5. ブートストラップ標本の *p* 値における累積確率分布を示す。
横に走る破線は，*p* = .01，*p* = .05 を示している。

3.3. ブートストラップ法を用いた効果量の信頼区間

ブートストラップ法はもちろん，効果量の信頼区間の推定にも援用することができる。ここでは，通例上 Cohen の *d* とよばれる効果量指標のひとつ，標準化平均差（水本・竹内，2011 に詳しい）の 95%信頼区間について検討する。例として 3.2 のデータを用いる。ブートストラップ標本のサイズを 22，*B* = 1,000 としてブートストラップ標本を生成し，それぞれについて標準化平均差を求める。その分布は図 6 のようになった。

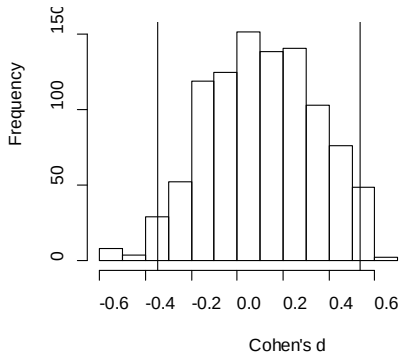

図 6. ブートストラップ標本の標準化平均差におけるヒストグラムを示す。
縦に走る実線は 95%信頼区間の下限および上限を表す。

ここからパーセンタイル法を用いて 95%信頼区間を求めると，[-0.34, 0.54]となった。やや正の方向に偏りがあるものの，信頼区間は 0 をまたいでいるため，効果量にもとづいて重要な議論をすることは勇み足になりかねないことがわかる。

また，前節と同様に累積確率分布を示すことも重要である（図 7）。たとえば，$d = 0$ となる累積確率をみることで，母集団から標本を抽出した際に，効果量が負の符号を取る確率などの見通しをつけることができる。ここではおよそ 40%弱程度であると考えられる。従来の仮説検定では，帰無仮説を積極的に採択することができないが，効果量の信頼区間は，母数が取りうる効果量の小ささという観点からもより自由に議論することを可能にする。

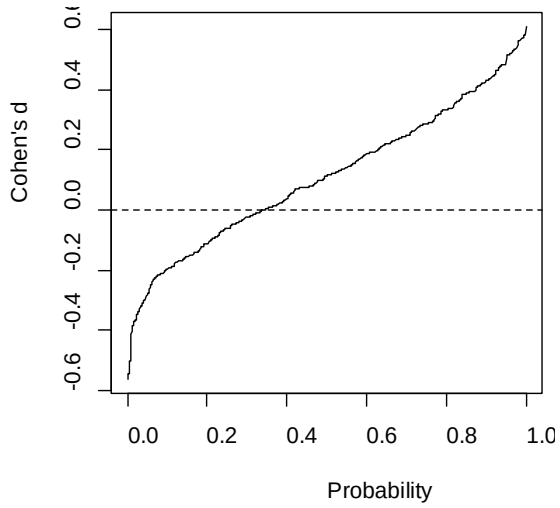

図 7. ブートストラップ標本の標準化平均差における累積確率分布を示す。
横に走る破線は，$d = 0$ を示している。

3.4. ブートストラップ法を用いたGroup Score Methodの誤差評価

ここではさらに，外国語教育研究に関連する典型的な分析手法を題材として，ブートストラップ法の応用可能性について見ていく。

外国語教育研究に関連が深い第二言語習得の分野では，70 年代，形態素習得順序研究（morpheme studies）が隆盛であった。形態素習得順序研究では，産出課題における義務的文脈（obligatory context）および正しく当該の形態素（ないし機能子）が用いられているかについての比率（正用率）を求め，それらの値についての項目間の順位を分析していた（e.g., Dulay & Burt, 1973）。分析に関してもさまざまな指標が開発され，それ自体が大きな問題であったという指摘もある（Rosansky, 1976）。最も代表的な指標は，Group Score Method（GSM）というもので，これはグループにおける義務的文脈の総数に対する正用数の比率（group score）で表される。

具体的な計算についての例を表 3 に示す。まず，実験協力者における課題中の正用数および義務的文脈について，形態素ごとに計算し，実験協力者全員について合計する。GSM では，グループに対してひとつの値が割り振られることになり，その値にもとづいて項目の順位づけをおこなう。このような順位が母語によるグループ間で一致する傾向が高いという研究が数多くなされた（e.g., Dulay & Burt, 1973）。

表 3
GSM における計算の例

実験協力者	進行形		過去形		冠詞	
	正用数	義務的文脈	正用数	義務的文脈	正用数	義務的文脈
A	3	4	13	40	50	140
B	3	6	14	43	32	103
C	3	4	14	65	31	180
D	1	2	5	40	120	201
合計	10	16	54	188	233	624
Group Score	.63		.29		.37	
順位	1 位		3 位		2 位	

形態素習得順序研究が第二言語習得理論に及ぼした影響ははかり知れず，いまもその多大な貢献については疑いようのないことである。しかしこの手法には，さまざまな問題点がある。まず，分母の異なる比率間の比較が分析上不都合となる場合が多い。次に，外れ値（ないし特定の個人的特性）の影響が group score に対して大きいと考えられることである。そして何より重要な点は，「ばらつき，標本誤差や信頼性を評価することができない」ことである。連続量（group score）を順位に変換することは多大な情報の損失である。

しかしブートストラップ法は，分布が未知であるような量についての推定精度を計算することを可能にする。たとえば，実験協力者についてノンパラメトリックブートストラップを用いて多数のブートストラップ標本を生成し，それらのブートストラップ標本毎に GSM を適用し，group score を得る。これらのブートストラップ標本の分布を検討することで，従来では得られなかった group score の推定精度の情報を得ることができる。

GSM に限らず，刺激再生法や思考発話法などに関わる発話データのコーディングスキーマにおいても同じような処遇が有効になる（e.g., Fukuta & Yamashita, 2014）。このような分析手法は外国語教育において一般的に用いられている。しかし誤差などの推定を経ぬまま積み重ねられてきた知見は，いうまでもなく再度見直すべきであろう。

3.5. その他の方法

その他にも，ブートストラップ法はさまざまな用途で用いられる。ブートストラップ法を用いた回帰分析，多項式曲線当てはめ（LOESS），クラスター分析（マルチスケールブートストラップ法，下平, 2002）などが広く知られている。また，判別分析や多変量解析でも用いられることもある。

4. ブートストラップ法と外国語教育研究

残念ながら，今日の外国語教育研究では，分析手法としてブートストラップ法が盛んに用いられているとはいいがたい。しかし，いくつか関連する研究は散見される。Larson-Hall and Herrington（2010）は頑健統計についての概説論文の中でブートストラップ法の紹介をしている。Plonsky et al.（2014）は *Language Learning*, *Studies in Second Language Acquisition* 掲載論文である 26 研究のデータを用いて，ブートストラップ法を用いた再分析をしている。草薙（2014b）は，外国語教育に関わる文処理研究の一部では，検定が対立仮説を採択できないとき，逆に帰無仮説を積極的に採択したような議論が多いことを指摘している。対立仮説を採択できないときは（そして理論的に帰無仮説の方に主眼があるときは），従来の確率分布を用いた通常の信頼区間推定に加え，ブートストラップ法を応用した平均値，平均差，効果量の信頼区間の推定，およびベイズ因子を用いた統計的処遇が有効であると述べている。Kusanagi（2014）では，英語の読解時間の比較の際に，ブートストラップ法を用いた効果量の信頼区間に基づいた議論をおこなっている。また，Fukuta and Yamashita（2014）は，発話データのコーディングに対してブートストラップ法を応用している。

母集団の分布が想定しにくい場合においても誤差などを推定することができる点が，ブートストラップ法の特色である。外国語教育研究では，一般的に現実的制約から小標本の研究が多く，さらに合理的にも明確な母集団を想定することができない場合がしばしばある（指導法研究などが特にそうである）。また，標本サイズの調整が困難であるため，検定力が十分に得られない場合も多い。さらに，他分野に比べ，種々の統計分析の前提を満たすことができない場合も見られる。外国語教育は，そもそも測定の妥当性を議論せずともよい工学や生物学などの分野に比べ，構成概念を研究対象としている。そのうえ，種々の現実的な側面によって測定の精度が全体的に低くなってしまう場合が多い。

このような中，頑健統計が担う分析上の役割は極めて大きいと考えられる（e.g., Larson-Hall, 2012）。特に測定の精度を推定する方法に幅を与えるブートストラップ法は，母分布が未知の場合でも，標本のみから測定の精度を導き出せるので，外国語教育研究との親和性もある程度見込まれる。

また，Plonsky et al.（2014）および Larson-Hall and Herrington（2010）では，第一種および第二種の過誤の管理についてもブートストラップ法の利点を述べている。標本の検定結果に依存するよりも，ブートストラップ法を用いた種々の検定が正しい意思決定を導く可能性もある。今後もさまざまな面でブートストラップ法を用いた分析がなされるだろう。

このようにブートストラップ法は頑健統計の手法のひとつとして，期待すべきものが大きいのであるが，もちろん万能ではない。ブートストラップ法はしばしば小標本を対象とするが，ブートストラップ法を用いたとしても小標本が母集団の代表性に対して問題を抱えていることには変わりない（under representation の問題）。Plonsky et al.（2014）が”To be sure, bootstrapping is not a replacement or cure-all for inadequate sampling”（p. 13）と述べているように，標本抽出自体に関わる問題をすべて賄うものではないのである。つまり元標本が不適切であれば，ブートストラップ法を用いた手法による結果も不適切になる場合が多い。このような点は，研究者が理論的見地や良識から判断するべきものであり，十分に注意すべきである。

結語となるが，どのような手法上の洗練をもってしても，それを用いる研究者の判断や解釈の重要性が失われることはない。むしろ，Plonsky et al.（2014）も述べるように，[3] 測定に関する構成概念とその妥当性を常に根幹に置くことが重要だと思われる。

本稿では，まずブートストラップ法の原理に関する基礎的な概説に努めた。特にブートストラップ法について見識を持たない研究者が理解しやすいように配慮したつもりである。次にブートストラップ法を用いた母数の区間推定などについて，いくつか事例をあげて紹介した。本稿が用いた分析にはすべて，R のデフォルト関数を用いて筆者が独自にプログラムしたものを用いている。現在ブートストラップ法を用いる分析手法は非常に多様なツールで実装されており，それらの使用方法について述べることは本稿の目的ではなかった。むしろ，実務的な使用方法や道具立てに拘泥せず，その背景となる基礎的な知識と考え方についての視野を提供することが本稿の目的である。本稿が今後の外国語教育研究におけるデータ分析手法の発展に貢献できるならば幸いこの上ない。

謝辞

本稿は，2014 年度外国語教育メディア学会関西支部メソドロジー研究部会第二回研究会で発表した内容にもとづいており，発表の際にフロアから頂いた質問の部分について加筆修正した部分がある。メソドロジー研究部会関係各位のみなさま，また，内容についてご教示を頂いた小林雄一郎先生（日本学術振興会）に深く感謝を申し上げる。

注

1. B = 1,000 という値に殊更意味はなく，計算資源が許す限り多ければ多いほどよいと考えられる。しかし通常 1,000 程度で十分であるといわれている。
2. これらのパッケージはほぼ同等の機能をもっている。
3. ここに Plonsky et al.（2014）の一節を引用する。この文章はブートストラップ法のみならず，外国語教育研究における分析手法全体にあてはまることである。

> “Finally and as always, no degree of statistical sophistication should ever take the place of principled analysis and interpretation based on an understanding of the data and the constructs they represent. It will always be important for researchers to take a step back from the statistical analysis

to evaluate the degree to which a particular technique is practically significant/useful in moving forward our knowledge of a given set of constructs."（Plonsky et al., 2014, p. 17）

引用文献

Dulay, H. and Burt, M. (1973). Should we teach children syntax? *Language Learning, 23*, 95–123.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall: New York.

Fukuta, J., & Yamashita, J. (2014). Interplay of Two Types of Cognitive Demands and Attention Orientation in L2 Oral Production. Paper presented at 17th World Congress of the International Association of Applied Linguistics (AILA), Brisbane, Australia

南風原朝和. (2002).『心理統計学の基礎―統合的理解のために』有斐閣アルマ.

草薙邦広. (2014a).「明示的および暗示的知識と学習者ビリーフの関係―英語を学習する日本の大学生を対象に―」『中部地区英語教育学会研究紀要』43, 87–92.

草薙邦広. (2014b).「英語の文法処理研究における統計的仮説検定：帰無仮説を主張する処遇について」『秋田英語英文学』55, 1–11.

草薙邦広. (2014c). 「外国語教育研究と直交表を用いた実験計画―実験計画の効率化を求めて―」『外国語教育メディア学会（LET）関西支部メソドロジー研究部会 2013 年度報告論集』24–33. Retrieved from http://www.mizumot.com/method/04-03_Kusanagi.pdf

草薙邦広. (2014d).「外国語教育研究における量的データの可視化―分析・発表・論文執筆のために―」『外国語教育メディア学会中部支部外国語教育基礎研究部会 2013 年度報告論集』53–70.

草薙邦広・水本篤・竹内理. (2014). 「日本の外国語教育研究における効果量・検定力・標本サイズ：Language Education & Technology 掲載論文を対象にした事例分析」『第 54 回外国語教育メディア学会全国研究大会発表要項』144–145.

Kusanagi, K. (2014). Processing flat adverbs in English as a foreign language: A preliminary self-paced reading study with highly proficient Japanese learners of English. *LET Journal of Central Japan, 25*, 63–72.

Larson-Hall, J. (2012). Our statistical intuitions may be misleading us: Why we need robust statistics. *Language Teaching, 45*, 460–474.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*, 368–390.

前田啓朗. (2000).「構成概念の妥当性の検証：日本の英語教育学研究における傾向と展望」『外国語教育評価学会研究紀要』3, 119–126.

水本篤・竹内理. (2008).「研究論文における効果量の報告のために―基礎的概念と注意点―」『関西英語教育学会紀要英語教育研究』31, 57–66.

水本篤・竹内理. (2011).「効果量と検定力分析入門―統計的検定を正しく使うために―」『より良い外国語教育研究のための方法：外国語教育メディア学会（LET）関西支部メソドロジー研究部会 2010 年度報告論集』47–73. Retrieved from http://www.mizumot.com/method/mizumoto-takeuchi.pdf

Mizumoto, A., Urano, K., & Maeda, H. (2014). A systematic review of published articles in *ARELE* 1–24: Focusing on their themes, methods, and outcomes. *ARELE*, *25*, 33–48.

下平英寿. (2002). 「ブートストラップ法によるクラスター分析のバラツキ評価」『統計数理』50, 33–44.

汪金芳. (2003).『計算統計 I − 確率計算の新しい手法（統計科学のフロンティア 11)』岩波書店.

大久保街亜・岡田謙介. (2012).『伝えるための心理統計: 効果量・信頼区間・検定力』勁草書房

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*, 450–470.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366.

Plonsky, L., Egbert, J., & LaFlair, G. T. (2014). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*. Advanced online publication. doi: 10.1093/applin/amu001

Wilcox, R. (2003). *Applying Contemporary Statistical Techniques*. Elsevier Science.

附録

R による処理の例

```
中央値のブートストラップ信頼区間

#変数の準備
bsmedians <- numeric(0)

#データ入力
dat <- c(8,5,3,4,5,3,8,9,14,6,7,7)

#ブートストラップ標本の生成（B = 1,000，サイズは 12，求める統計量は中央値）
for(i in 1:1000){
bsamples <- sample(dat, 12, replace = T)
bsmedians[i] <- median(bsamples)
}

#信頼区間の算出（パーセンタイル法，95%）
quantile(bsmedians, c(.025, .975))
```