〔東京家政大学研究紀要　第48集 (1), 2008, pp.1～7〕

ブートストラップBCa法による効果量δの信頼区間

井上 俊哉

(平成19年10月4日受理)

Bootstrap BCa Confidence Intervals for Effect Size δ

INOUE, Shunya

(Received on October 4, 2007)

キーワード：効果量，標準化平均値差，信頼区間，ブートストラップBCa法，シミュレーション
Key words : effect size, standardized mean difference, confidential interval, Bootstrap BCa method, simulation

1 はじめに

1.1 仮説検定の偏重

心理学研究において，検定は欠かせない手法となっている．1992年から1993年の間に，心理学研究，教育心理学研究，社会心理学研究など，心理学関係の学術雑誌に掲載された論文を調べた尾見・川野(1994)[1]によると，7誌に掲載された全256論文中54.3%に当たる139論文で分散分析，40.6%に当たる104論文でt検定，20.3%に当たる52論文でカイ2乗検定が用いられており，検定が全く用いられていない研究は54論文(21.1%)だったという．2005年から2006年の間の教育心理学研究掲載論文を調べた栗田(2007)[2]でも，22%で分散分析，13%で相関係数の検定，8%でt検定，6%でカイ2乗検定が用いられていたことが示されている．心理学科の学生向けの統計学の授業・教科書においても，検定の占める比重は非常に大きい．

このように，検定は心理学研究における方法論として不可欠ともいえる中心的な地位を占めているが，その偏重・誤用に対しては，古くから批判も繰り返されてきた(Carver,1978[3]；Cohen, 1994[4]；Rozeboom, 1960[5]；Oakes, 1986[6]；橘(1986)[7]など)．たとえば，Oakes (1986)は，(1)仮説検定の結論が2分法であること，(2)帰無仮説と対立仮説の関係が対称的ではないこと，(3)サンプルサイズを大きくすればどんな仮説も棄却しうること，(4)検定の結論が有意であったとしても，差(関連)の大きさの程度を示すことができないことを指摘している．

1.2 効果量と信頼区間

検定の偏重を批判する論者の多くは，検定の限界を克服するために効果量とその信頼区間を報告することを提唱しており(Cohen,1994[4]；Oakes,1986[6]；Schmidt, 1996[8]；Wilkinson, et al.[9]など)，アメリカ心理学会の *Publication Manual* でも効果量，信頼区間の報告が強く推奨されるに至っている(American Psychological Association, 2001[10], p.22)．仮説検定のかわりに(あるいは仮説検定に加えて)効果量と信頼区間を報告することによって，以下のような効用が得られると考えられる．

(1) 仮説検定の結果，群間差(変数間の関連)が有意だと確認されても，差(関連)が「ある」ことが示されるだけで，差や関連の程度・強さは明示されない．p値が0に近いことをもって差や関連が大きいとする解釈も見られるが，微少な差や関連しかなくてもサンプルサイズが大きくなることでp値は0に近づく．これに対して，効果量は差や関連の強さを直接的に示すものである．

(2) 仮説検定の結論は2分法であるため，ほんのわずかなp値の違い(たとえば，p=0.048とp=0.051)により，一方は有意差あり，他方は有意差なしというまったく逆の結論を下さざるを得ない．結論の正しさに関しても，正しいか誤っているかの2分法である．これに対して，効果量と信頼区間を用いれば，差(関連)の大きさを連続的に評価することができる．しかも，構成された信頼区間が0を含むかどうかに目を向ければ，検定の結論も同時に得ることができ

教養部情報処理研究室

る．

(3) 仮説検定に際して検定力に注意が向けられることがほとんどなく(Cohen, 1962[11]；Sedlmeier & Gigerenzer, 1989[12])，サンプルサイズが不適切なまま研究が行われることが少なくないと考えられる．これに対して，効果量推定値の位置と区間の幅は，適切なサンプルサイズを決めるための重要な情報を提供する．

(4) 仮説検定における有意水準は，検定の結論が誤っている確率を表すものではない．それに対して信頼区間の信頼係数は，区間が正しく母数を含む確率を示しており，解釈が自然である．

(5) 研究結果の蓄積という観点からも，効果量と信頼区間の利用が勧められる．複数の研究が報告する検定統計量の値（Fやtなど）と有意水準を集めても正しい結論に近づくことはできない．これに対して，複数の研究から得られた効果量推定値の全体的な傾向，区間の重なり具合などは，真の効果量へ近づく有益な手がかりとなる．また，報告された効果量の値は，そのままメタ分析に活用できる．

研究報告において効果量と信頼区間を示すことは，上記のように多くの長所を持つと考えられる．しかし，米国においてもこれらを報告する慣行は常識になっているとはいえず，日本ではこれらを報告しようという機運すらない．Steiger and Fouladi(1997)[13]は，心理学研究で区間推定があまり用いられない理由として，(1)仮説検定が好まれ，区間推定が行われてこなかった(*Tradition*)；(2)帰無仮説の棄却を目標とする状況で信頼区間を用いてもあまり有用でない(*Pragmatism*)；(3)標準的なテキストで信頼区間が論じられず，多くの心理学者は信頼区間に興味がない(*Ignorance*)；(4)多くのパラメータに関して，信頼区間の算出はコンピュータの利用が前提となるが，主要な統計パッケージは区間推定に対応していない(*Lack of availability*)という4点を挙げている．Steiger and Fouladiの指摘する状況を一朝一夕に変えることは難しいが，信頼区間を「使いたいのに使えない」状況は変わる必要があるだろう．SPSS, SAS, S-plusで効果量の信頼区間を求めるスクリプト(Smithson, 2003[14])がインターネット上に公開されているものの，日本の一般研究者が気軽に区間推定を行う環境が整っているとは言い難い．以下では，種々の効果量指標の中でももっとも代表的な指標δ（デルタ）の信頼区間構成法を概観したのち，Excel VBAを用いて作成したδの信頼区間構成マクロについて報告する．

2. 効果量δの信頼区間

2.1 効果量δの点推定

δは，2群の母集団平均の差（$\mu_1-\mu_2$）を両群に共通の標準偏差（σ）で割って標準化したもので，標準化平均値差とも呼ばれ，メタ分析における重要な指標としても知られている．

$$\delta=\frac{\mu_1-\mu_2}{\sigma}$$

（δについて「Cohenのd」と表記されることも多いが，母集団のパラメータはギリシャ文字で表記するという通常の統計学の慣習およびHedges & Olkin, 1985[15]の表記にしたがって，ここではδの表記を用いる）．

δの推定量としてもっともよく用いられるのは，2群の標本平均の差を両群に共通な標準偏差の推定値

$s=\sqrt{\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ で割ったHedgesのgである（n_1, n_2は各群のサンプルサイズ，s_1^2, s_2^2は各群の不偏分散）．

$$g=\frac{\bar{x}_1-\bar{x}_2}{s}$$

gはδの不偏推定量ではないため，偏りを修正した推定量としてdが用いられることもある（Γ(x)はガンマ関数）．

$$d=\frac{\Gamma\left(\dfrac{n_1+n_2-2}{2}\right)}{\left(\dfrac{n_1+n_2-2}{2}\right)\Gamma\left(\dfrac{n_1+n_2-3}{2}\right)}\times g$$

gに乗じられる定数部分は，サンプルサイズが大きくなると1に近づくため，サンプルサイズが大きいときにはgとdの違いは小さい．たとえば，$n_1=n_2=10$のときには，$d\fallingdotseq 0.9576\times g$だが，$n_1=n_2=100$になると，$d\fallingdotseq 0.9962\times g$である．

2.2 効果量δの区間推定

2.2.1 近似式に基づく方法

母集団分布に正規分布を仮定すると，サンプルサイズが大きくなるとき，dの分布は近似的に，平均δ，分散 $\sigma_d^2=\dfrac{n_1+n_2}{n_1n_2}+\dfrac{\delta^2}{2(n_1+n_2-2)}$ の正規分布にしたがう．

σ_d^2の推定量として $\hat{\sigma}_d^2 = \frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2-2)}$ を用いると，δに関する95%信頼区間の下限は $\delta_L = d - z_{1-\alpha/2} \times \hat{\sigma}_d^2$，上限は $\delta_U = d + z_{1-\alpha/2} \times \hat{\sigma}_d^2$ で求めることができる．ただし，$z_{1-\alpha/2}$ は標準正規分布における下側確率 $1-\alpha/2$ に対応する値である．δが小さく，サンプルサイズがある程度大きいとき，この近似はかなり正確である（Hedges & Olkin, 1985[15]）．

2.2.2 変数変換による方法

フィッシャーのz変換を用いて母相関係数の信頼区間を求めるのと同様に，標本効果量dを変換して，より単純な分布で信頼区間を求めた上で，得られた区間をδの区間に再変換するという方法もある．

標本効果量dを $h(d) = \sqrt{2}\sinh^{-1}\frac{d}{a} = \sqrt{2}\log\left(\frac{d}{a} + \sqrt{\frac{d^2}{a^2}+1}\right)$

によって変換したものをh，母集団効果量δを同じ変換で変換したものをηとすると，$\sqrt{n_1+n_2}(h-\eta)$ は標準正規分布に近似する（ただし，$a = \sqrt{4+2(n_1/n_2)+2(n_2/n_1)}$）．このことから，$\eta$に関する信頼区間の下限は $\eta_L = h - z_{1-\alpha/2}/\sqrt{n_1+n_2}$，上限は $\eta_U = h + z_{1-\alpha/2}/\sqrt{n_1+n_2}$ で求められる．これを逆変換することで，δに関する信頼区間の下限 $\delta_L = h^{-1}(\eta_L)$ と上限 $\delta_U = h^{-1}(\eta_U)$ を得ることができる．

2.2.3 非心t分布に基づく方法

δの区間推定法としてもっとも代表的なのは，非心t分布に基づく方法である（Steiger & Fouladi, 1997[13]; Cumming & Finch, 2001[16]）．2群の平均値差を標準誤差で割ったtは，2群の母集団平均が等しいという仮説のもとでは自由度n_1+n_2-2のt分布にしたがうが，仮説が正しくないときには，自由度n_1+n_2-2，非心パラメータλの非対称な非心t分布にしたがう．また，非心パラメータλと効果量δの間には，$\lambda = \delta\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ という関係が成り立つ．δの区間推定のためには，非心パラメータλに関する信頼区間をはじめに構成し，λとδの関係式に基づき，δの信頼区間を構成するという手順がとられる．

2.2.4 ブートストラップ法

非心t分布を利用した区間推定は，正規性，等分散などを仮定して行われるが，この種の仮定を必要としない区間推定法として注目されるのがブートストラップ法である（Efron,B., & Tibshirani,R.J., 1993[17]；汪・田栗, 2003[18]）．ブートストラップ法に基づく信頼区間構成法として，パーセンタイル法，BCa法，ブートストラップt法などがある．このうち，パーセンタイル法は精度が低く，t法は理論的には優れているものの実際にはうまく機能しないことが多いとされる．BCa法は以下のようなステップをたどる（BCaは，Bias-Corrected and acceleratedの略）．

1)手元のデータからの独立な復元抽出によって，ブートストラップ標本を構成する（サンプルサイズは手元のデータと同じ）．

2)ブートストラップ標本に基づき，効果量δの推定値を求める．

1)2)のステップを，B回繰り返す．

3)1)2)のステップを通じて，加速定数$\hat{a}$と偏り修正値$\hat{z}_0$を求める．

4)信頼区間の下限に対応するパーセンタイル点CI_Lと上限に対応するパーセンタイル点CI_Uのそれぞれを，以下の式により求める（式中のΦは標準正規分布の分布関数である）．

$$CI_L = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}\left(\hat{z}_0+z_{\alpha/2}\right)}\right) \quad CI_U = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{a}\left(\hat{z}_0+z_{1-\alpha/2}\right)}\right)$$

5)B回分のブートストラップサンプルから計算されたB個の推定量の分布において，CI_Lに対応する値を信頼区間の下限，CI_Uに対応する値を信頼区間の上限とする．

ブートストラップBCa法は母集団の形状に関する仮定を必要としない点で，現実場面での利用価値が高いと考えられる．本研究では，ブートストラップBCa法による信頼区間を算出するExcelマクロを作成した．

3. 効果量δの信頼区間を構成するためのExcelマクロ

3.1 Excelマクロ

ブートストラップBCa法のアルゴリズムに関する汪・田栗(2003)[18]の解説，Kelley(2005)[19]のS-plus用シンタックスを参考に，Excel VBAを用いてマクロを作成した．Excelシート上のデータが入力されている範囲の左上隅セルと，結

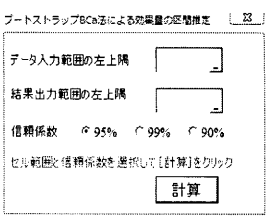

図1 ブートストラップBCa法Excelマクロ

果を出力させたい範囲の左上隅セル，および信頼係数を選択すると，効果量の点推定値と信頼区間の下限と上限が算出され，出力されるように作ってある．効果量δの推定量としては，gではなくdを用いている（図1）．

3.2 シミュレーション

ブートストラップBCa法によるδの信頼区間に関するシミュレーション研究として，Kelley(2005)[19]とAlgina et al.(2006)[20]がある．Kelley(2005)は，「母集団分布が正規分布でδ=0」「母集団分布が非正規分布でδ=0」「母集団分布が非正規分布でδ≠0」という3条件について，非心t分布による信頼区間(NCT)，ブートストラップ・パーセンタイル法による信頼区間，ブートストラップBCa法による信頼区間(BCa)の3種類の方法で構成された信頼区間を比較している．その結果，パーセンタイル法による信頼区間は明らかに不正確であり劣っていること，正規・非正規を問わず，δ=0の場合には，NCT，BCa法によって構成された信頼区間が真のδを含む割合は名目上の信頼係数に非常に近いことを示している．ただし，サンプルサイズが小さいときにはBCa法の構成する信頼区間の幅はやや広くなる傾向があった．Kelleyは，非正規分布についてδ≠0のシミュレーションを行っているが，δ=0.2，δ=0.5，δ=0.8，δ=1.6の各効果量について，それぞれ一通りのサンプルサイズしか設定しておらず（検定力80%になるようにサンプルサイズを決めているため），シミュレーション結果に影響しているのが効果量なのかサンプルサイズなのかが不分明であることが難点となっている．

Algina et al.(2006)のシミュレーションは，Kelleyよりも極端な非正規条件を設定しているほか，非正規かつδが0でないケースについて，Kelleyよりも細かい状況分けをして検証している．その結果，正規分布からの逸脱が極端で，しかもδが大きいケースでは，NCTによる信頼区間が不正確であること，BCa法もNCTほどではないが信頼係数の正確さが損なわれることが示されている．

3.2.1 両群とも正規分布の場合

本研究ではまず正規性，等分散，独立の条件が満たされている場合について，95%信頼区間を10,000回繰り返し求めるシミュレーションを行った．正規乱数生成法は，縄田(2003)[21]によった．ブートストラップ・サンプリングの回数Bについて，Efron and Tibshirani(1993)[17]には少なくとも1,000以上とあるが，ここではKelley(2005)にならい10,000に設定した．サンプルサイズは各群15，50，75の3通り，効果量δは0.0，0.2，0.5，0.8，1.4の5通りを設定した．なお，0.2は小さな効果量，0.5は中程度の効果量，0.8は大きな効果量の目安とされる値であり（Cohen,1988[22]），また，実際の心理学研究において観測される効果量の分布の中央値は約0.5になるという報告もある（Lipsey & Wilson,1993[23]）．

シミュレーションの結果を表1にまとめた．表中の% of Coverageは「構成された信頼区間が設定したδを含む割合」，Mean widthは「構成された信頼区間の幅の平均」，Mean unbiased dは「推定された不偏dの平均」である．表1より，ここで試したすべての状況において，構成された信頼区間が設定したδを含む割合は，名目上の信頼係数95%に非常に近い値になっていることが確認できる．あえていえば，$n_1=n_2=15$，δ=1.4のケースにおける% of Coverageがいくらか小さい(0.9339)が，これもAlgina et al.(2006)が許容範囲として採用した0.925～0.975には十分おさまっており，両群の母集団分布がともに正規分布である場合，ブートストラップBCa法は非常に正確な信頼区間を構成するといえる．

表1　2群とも正規分布（B=10,000）

sample size		effect size δ				
		0.0	0.2	0.5	0.8	1.4
$n_1=n_2=15$	% of Coverage	0.9482	0.9452	0.9446	0.9419	0.9339
	Mean width	1.4834	1.4874	1.5024	1.5279	1.6172
	Mean unbiased d	0.0067	0.2025	0.4942	0.8004	1.3986
$n_1=n_2=50$	% of Coverage	0.9483	0.9524	0.9510	0.9457	0.9440
	Mean width	0.7907	0.7929	0.8021	0.8190	0.8759
	Mean unbiased d	0.0016	0.1986	0.4971	0.8003	1.3956
$n_1=n_2=75$	% of Coverage	0.9500	0.9531	0.9490	0.9480	0.9457
	Mean width	0.6435	0.6452	0.6528	0.6673	0.7138
	Mean unbiased d	-0.0019	0.1990	0.4974	0.8020	1.3987

表2　2群とも歪度1.75，尖度6.75（B=10,000）

sample size		effect size δ				
		0.0	0.2	0.5	0.8	1.4
$n_1=n_2=15$	% of Coverage	0.9411	0.9358	0.9299	0.9139	0.8807
	Mean width	1.4484	1.4571	1.4999	1.5781	1.8029
	Mean unbiased d	-0.0003	0.2155	0.5305	0.8408	1.4618
$n_1=n_2=50$	% of Coverage	0.9464	0.9516	0.9439	0.9368	0.9289
	Mean width	0.7817	0.7863	0.8110	0.8551	0.9896
	Mean unbiased d	0.0043	0.1977	0.5070	0.8120	1.4200
$n_1=n_2=75$	% of Coverage	0.9501	0.9510	0.9482	0.9381	0.9331
	Mean width	0.6381	0.6420	0.6634	0.7002	0.8142
	Mean unbiased d	-0.0010	0.2012	0.5037	0.8070	1.4145

3.2.2 両群とも歪度1.75，尖度6.75の場合

非正規分布のありようは無数に存在し，すべての非正規分布について網羅的に調べることは不可能である．ここでは，Kelley(2005)が調べたもっとも極端なケースである両群とも歪度1.75，尖度6.75というケースについてのみ検討した．Kelley(2005)では，δ≠0の場合について，各δでサンプルサイズが1種類しか設定されていな

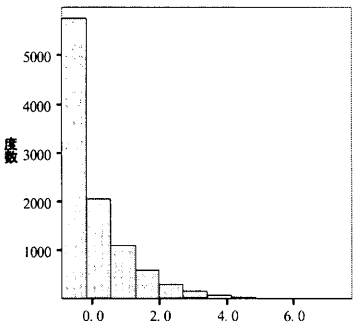

図2　歪度1.75，尖度6.75で生成されたデータ

かったが，ここでは，両群とも正規分布の場合と同様に，δの値ごとに3種類のサンプルサイズを設定してシミュレーションを行った．歪度1.75，尖度6.75の分布を生成するのには，Kelleyと同様，Fleishman(1978)[24]のpower methodを用いた．図2は，10,000個のデータを生成して描いたヒストグラムである．

シミュレーションの結果は表2の通りである．明らかな傾向として認められるのは，効果量δが大きいほど，サンプルサイズが小さいほど，BCa法によって構成された信頼区間がδを含む割合が小さくなっていることである．効果量δが1.4と大きく，各群のサンプルサイズが15と小さいときには，信頼区間がδを含む割合は0.8807で，さすがに正確とはいえない値である．この知見はAlgina et al.(2006)[20]とも一致する．正規分布を仮定するNCTと比べると正確であるとはいえ，効果量大，サンプルサイズ小という条件下での区間推定には慎重である必要があろう．逆に，効果量δが小さいときには(たとえば，歪度1.75，尖度6.75という本シミュレーションの状況下ではδ=0.5以下)，各群15名程度の小さなサンプルサイズでもかなり正確な信頼区間が得られている．

3.2.3 ブートストラップ・サンプリングの回数Bを変えることの影響

ブートストラップ・サンプリングの回数Bについて，Efron and Tibshirani(1993)[17]は，正確な推定のためには少なくとも1,000以上としている．今回のシミュレーションでは慎重を期して，Kelley(2005)にならいB=10,000に設定したが，Algina et al.(2006)はB=1,000を採用している．B=1,000でも同程度に正確な信頼区間が得られるならば，コストの削減につながる．そこで，表1および表2で示したのと同じシミュレーションを，B=1,000で行った．両群正規分布の結果が表3，両群とも歪度1.75，尖度6.75での結果が表4である．

表3　2群とも正規分布（B=1,000）

sample size		effect size δ				
		0.0	0.2	0.5	0.8	1.4
$n_1=n_2=15$	% of Coverage	0.9482	0.9430	0.9414	0.9408	0.9337
	Mean width	1.4759	1.4811	1.4968	1.5239	1.6090
	Mean unbiased d	0.0004	0.1999	0.5027	0.7986	1.3987
$n_1=n_2=50$	% of Coverage	0.9493	0.9494	0.9525	0.9463	0.9402
	Mean width	0.7873	0.7895	0.7987	0.8158	0.8700
	Mean unbiased d	0.0002	0.2011	0.4991	0.7990	1.4006
$n_1=n_2=75$	% of Coverage	0.9470	0.9523	0.9483	0.9441	0.9421
	Mean width	0.6411	0.6422	0.6500	0.6644	0.7099
	Mean unbiased d	0.0018	0.2004	0.4984	0.8005	1.3968

表4　2群とも歪度1.75、尖度6.75（B=1,000）

sample size		effect size δ				
		0.0	0.2	0.5	0.8	1.4
$n_1=n_2=15$	% of Coverage	0.9364	0.9361	0.9319	0.9122	0.8813
	Mean width	1.4416	1.4504	1.5023	1.5689	1.7953
	Mean unbiased d	0.0008	0.2132	0.5202	0.8404	1.4676
$n_1=n_2=50$	% of Coverage	0.9449	0.9460	0.9381	0.9336	0.9178
	Mean width	0.7783	0.7829	0.8076	0.8506	0.9843
	Mean unbiased d	-0.0011	0.2015	0.5066	0.8093	1.4248
$n_1=n_2=75$	% of Coverage	0.9486	0.9484	0.9410	0.9399	0.9270
	Mean width	0.6351	0.6395	0.6597	0.6970	0.8062
	Mean unbiased d	-0.0007	0.2027	0.5052	0.8087	1.4155

表1と表3を比較すると，2群とも正規分布の場合には，B=1,000でもほぼ遜色のない結果が得られていることがわかる．2群の母集団分布がいずれも歪度1.75と尖度6.75である表2と表4を比べると，$n_1=n_2=50$，$n_1=n_2=75$の場合には，一部の例外を除き，B=10,000の% of Coverageが名目上の信頼係数0.95に若干近いが，サンプルサイズが小さい（$n_1=n_2=15$）ケースでは，B=10,000にしてもあまり改善がみられない．Bをさらに大きくとることで信頼区間を正確にできるかどうかをみるために，「$n_1=n_2=15$，$\delta=0.8$」「$n_1=n_2=15$，$\delta=1.4$」「$n_1=n_2=50$，$\delta=1.4$」の3条件についてのみB=15,000にしてシミュレーションを試みた．その結果，3条件における% of Coverageのみ順に記すと，0.9138，0.8849，0.9291であり，ほとんど改善はみられなかった．

以上を総合すると，母集団分布が正規分布に近い場合にはB=1,000でも十分であること，母集団分布が正規分布から逸脱している場合，サンプルサイズが大きければB=10,000にとることで若干の改善が望めることがわかる．分布が非正規で，δが大きい場合に信頼区間を求めるには，サンプルサイズを大きくとることが何よりも大事である．

3.2.4 両群のサンプルサイズが等しくない場合

最後に，2群のサンプルサイズの偏りの影響をシミュレーションによって検討した．母集団分布の平均は，第1群がδ，第2群が0，標準偏差，歪度，尖度は2群に共通で，それぞれ，1.00，1.75，6.75とした．各群のサン

表5　2群とも歪度1.75、尖度6.75（B=1,000）

sample size		effect size δ				
		0.0	0.2	0.5	0.8	1.4
$n_1=n_2=25$	% of Coverage	0.9393	0.9367	0.9347	0.9241	0.9032
	Mean width	1.1036	1.1109	1.1444	1.2044	1.3771
	Mean unbiased d	-0.0046	0.2059	0.5113	0.8220	1.4360
$n_1=n_2=50$	% of Coverage	0.9449	0.9460	0.9381	0.9336	0.9178
	Mean width	0.7783	0.7829	0.8076	0.8506	0.9843
	Mean unbiased d	-0.0011	0.2015	0.5066	0.8093	1.4248
$n_1=25, n_2=75$	% of Coverage	0.9419	0.9429	0.9374	0.9386	0.9308
	Mean width	0.8871	0.9001	0.9279	0.9701	1.0951
	Mean unbiased d	-0.0016	0.2017	0.5123	0.8140	1.4174
$n_1=75, n_2=25$	% of Coverage	0.9398	0.9297	0.9236	0.9178	0.9144
	Mean width	0.8935	0.8845	0.9026	0.9408	1.0548
	Mean unbiased d	-0.0043	0.2019	0.5130	0.8111	1.4239
$n_1=n_2=75$	% of Coverage	0.9486	0.9484	0.9410	0.9399	0.9270
	Mean width	0.6351	0.6395	0.6597	0.6970	0.8062
	Mean unbiased d	-0.0007	0.2027	0.5052	0.8087	1.4155

プルサイズは，$n_1=75$，$n_2=25$の条件と$n_1=25$，$n_1=75$の条件を用意した．いずれの条件についてもB=1,000とした．シミュレーションの結果は表5の通りである．平均が大きい方の群（第1群）のサンプルサイズが大きいときに，信頼区間が効果量を含む割合の正確さが損なわれることを見て取ることができる．この一例から過度な一般化をすることはできないが，できるだけ2群のサンプルサイズを揃える方が効率的といえそうである．

4 まとめ

効果量と信頼区間は，日本では注目されることが少ないが，検定では得られない有益な情報をもたらしてくれることは間違いない．だが，統計を専門としない一般の心理学研究者が効果量や信頼区間を用いるためには，統計教育やソフトウェアなど，さまざまな面からの環境整備が必要である．また，効果量の利用が普及しない原因の一つとして，効果量の種類が多様であること，効果量指標の性質についての知識が普及していないことが挙げられる．本研究では種々ある効果量指標のうち，もっとも用いられることの多い基本的な指標である独立な2群の標準化平均値差δに注目した．まずは，このδについて，一般心理学研究者の認知度が上がり，実際の研究で点推定，区間推定を報告する研究者が増えることが望まれる．そこで，本研究では，Excel上で簡単に標準化平均値差δの点推定値と信頼区間を求めることのできるマクロを作成した．標準化平均値差δの信頼区間構成法としては，非心t分布に基づく方法が代表的であるが，ここでは，母集団分布の形状を前提しないブートストラップBCa法を採用した．シミュレーションの結果，母集団分布が正規分布に近い状況ではサンプルサイズがそれほど大きくなくても（各群15名），マクロがかなり正確な信頼区間を構成することが確認された．母集団分布が正規分布でない状況は無数にあるため，非正規のケースについて一般的な結論を導くこはできないが，シミュレーションの結果，母集団分布が正規性から大きく逸脱している場合，δが大きくサンプルサイズが小さいときには，構成される信頼区間の正確さが損なわれることが示唆された．母集団分布の形状が正規分布と大きく食い違い，δが大きいと思われる状況では，大きなサンプルサイズを確保する必要があるだろう．ブートストラップ法のブートストラップ・サンプリング回数について，どの程度の繰り返しが必要かについて，シミュレーションにより調べたところ，B=1,000程度をとれば十分であるようだった．

引用・参考文献

[1] 尾見康博・川野健治 (1994) 心理学における統計手法再考－数字に対する"期待"と"不安"－ 性格心理学研究,2(1),56-67.

[2] 栗田佳代子 (2007) 測定・評価に関する研究動向と展望－統計的データ解析法の利用の現状とこれから－ 教育心理学年報第46集，102-110.

[3] Carver,R.P.(1978). The case against statistical significance testing. *Harvard Educational Review, 48, 378-399.*

[4] Cohen,J. (1994). The earth is round (p<.05). *American Psychologist,* 49 (12), 997-1003.

[5] Rozeboom,W.W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin,* 57, 416-428.

[6] Oakes,M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* John Wiley & Sons.

[7] 橘敏明 (1986) 医学・教育学・心理学にみられる統計的検定の誤用と弊害. 医療図書出版社

[8] Schmidt,F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods,* 1 (2), 115-129.

[9] Wilkinson,L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journal: Guidelines and explanations. *American Psychologist,* 54 (8), 594-604.

[10] American Psychological Association (2001). *Publication manual of the American Psychological Association.* Washington, DC: American Psychological Association.

[11] Cohen,J. (1962). The statistical power of abnormal-social psychology research:A review. *Journal of Abnormal and Social Psychology,* 65, 145-153.

[12] Sedlmeier,P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin,* 105, 309-316.

[13] Steiger ,J.H., & Fouladi, R.T. (1997). No centrality interval estimation and the evaluation of statistical models. In L.L.Harlow, S.A.Mulaik, & J.H.Steiger (Eds.), *What if there were no significance tests?* (pp.221-257). Mahwah, NJ: Lawrence Erlbaum Associates.

[14] Smithson,M.J. Scripts and Software for Noncentral Confidence Interval and Power Calculations [Computer software]. Retrieved July 7, 2007 http://psychology.anu.edu.au/people/smithson/details/CIstuff/CI.html

[15] Hedges,L.V., & Olkin, I. (1985). *Statistical Methods for meta-analysis.* Orlando, FL: Academic Press.

[16] Cumming,G., Finch,S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distribution. *Educational and Psychological Measurement,* 61, 532-574.

[17] Efron,B., & Tibshirani,R.J. (1993). *An introduction to the bootstrap.* New York:Chapman & Hall.

[18] 汪金芳,田栗正章 (2003) ブートストラップ法の基礎 甘利俊一・竹内啓・竹村彰通・伊庭幸人（編）統計科学のフロンティア第11 巻：計算統計 I - 確率計算の新しい手法 (pp.1-64) 岩波書店.

[19] Kelley,K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement,* 65 (1), 51-69.

[20] Algina, J., Keselman, H.J., & Penfield, R.D. (2006) Confidence interval coverage for Cohen's effect size statistics. *Educational and Psychological Measurement,* 66 (6), 945-960.

[21] 縄田和満 (2003) Excelによる確率入門 朝倉書店

[22] Cohen,J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale,NJ: Erlbaum.

[23] Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist,* 48, 1181-1209.

[24] Fleishman,A.I. (1978). A method for simulating non-normal distribution. *Psychometrika,* 43(4), 521-532.

Abstract

In psychological research it seems that hypothesis significance testing has been given too much emphasis; it is time for researchers to reconsider the importance of reporting confidence intervals and effect sizes. Various factors may hinder this kind of reporting, and one of the reasons could be that there is currently no software available for calculating these statistics easily; as we know it is practically impossible to calculate the values without computers. Recently, bootstrap methods have been attracting attention for their robustness and no need to make any particular assumption about the distribution of a population. In this study, we developed an Excel macro computer program to estimate confidence intervals around the standardized mean difference δ based on the bootstrap BCa procedure. The results of our simulation show that our program can precisely estimate confidence intervals except when all the following situations happen simultaneously: the normality assumption is extremely violated, effect size is large, and sample size is very small.