『より良い外国語教育研究のための方法』(pp. 47–73)
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

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―

水本 篤
関西大学

竹内 理
関西大学

キーワード: 統計的検定，有意差，効果量，検定力，検定力分析

1. 本稿の目的

統計的検定は，標本から得たデータ分析結果を母集団にまで一般化させる目的で行われる。統計的検定では，サンプル・サイズ，有意水準，効果量，検定力の4つが検定結果の良し悪しを決定する重要な要素であるため，その基礎的概念の理解が検定を正しく使うためには重要である。そこで，本稿では，効果量と検定力分析の2つの概説を行い，統計的検定を用いている研究において，効果量報告と検定力分析の使用を推奨することを目的とする。

2. 効果量

2.1 統計的検定と効果量

統計的検定では，たとえば，手元のデータ（標本，sample）である 2 つのグループの平均値に差がありそうだと考える場合に，母集団（population）でもその平均値差は同じように見られるであろうかということを推定する。その際に，「平均値には差がない」という，主張したいこととは逆の仮説をとりあえず立てて，その「平均値の差がない」確率が低い場合には，「平均値差がある」と判断するという論法になっている。その確率は英語の probability から，p 値と呼ばれており，手元のデータから計算することが可能である。

p 値がどれくらい小さければ統計的に有意な差があるかは，データ収集よりも前に設定する有意水準（significance level）に基づいて判断される。慣例として，有意水準は 5%（分野や研究内容によっては 1%）に定められている。[1] そのため，収集したデータに対して検定を行い，結果が $p < .05$ であれば「差がある」と判断される。統計的検定は基本的に「差がある」と主張するために行われるため，$p < .05$ であれば，望ましい結果が得られたと結論づける。通常，どの分野のジャーナルでも，有意な結果（$p < .05$）が出た研究論文を掲載する傾向があるため，p 値が統計的検定においてもっとも重要な指標であると考えてしまう。

しかし，p 値はサンプル・サイズ（標本数，サンプル数）が大きくなればなるほど，実質的な差がなかった場合でも，p 値は小さくなり，統計的に有意であるという結果が得られやすくなるという大きな問題を持っている。そのため，ある検定を行ったところ，50 人では有意ではなく，100 人のデータの場合には有意になるということも十分にあり得る。その具体例としてシミュレーションによるデータを表 1 に示す。データセット A，データセット B ともに，グループ 1 は平均値 40，標準偏差 10，グループ 2 は平均値 43，標準偏差 10 となるように設定して，正規分布の乱数を発生させた。[2] 平均値の差は両データセットとも 3.0 である。これら 2 つのデータセットに対して，対応のない t 検定（independent t-test）を行った結果，データセット A では $p = 0.137$（$p > .05$），データセット B では $p = 0.015$（$p < .05$）という結果となる（つまり，人数の違いによって，有意差のあり・なしが変わっている）。この例は，同じ平均値，標準偏差であっても，サンプル・サイズが大きくなればなるほど p 値が小さくなり，「有意差あり」という判断をしやすくなるということを示している。また，よくある誤った解釈である「p 値が小さければ小さいほど，差が大きい」というものが間違いであるということもわかる。

表 1

シミュレーション・データでの n の違いによる p 値の比較

データセット	グループ	n	平均値	標準偏差	平均値差	対応のない t 検定	効果量 d
A	グループ 1	50	40	10	3.0	$p = 0.137$	$d = 0.3$
	グループ 2	50	43	10			
B	グループ 1	100	40	10	3.0	$p = 0.015$	$d = 0.3$
	グループ 2	100	43	10			

効果量 d の基準：$d = 0.2$（効果量小），$d = 0.5$（効果量中），$d = 0.8$（効果量大）

上述のように，p 値はサンプル・サイズによって変わるものなので，実質的な差が大きいか小さいかについての情報は何も与えてくれない。そこで，サンプル・サイズによって変化しない，標準化された指標である効果量（effect size）が解釈に役立つ。「グループごとの平均値の差を標準化した効果量」の代表的な指標である Cohen's d は，t 検定のような 2 グループの平均値の差を比較するときに使用し，平均値の差の効果量を以下のような式 (1) で求めることができる（実験群と統制群のサンプル・サイズが同じ場合）。[3] 一見，難しそうに見えるかもしれないが，実は平均値と標準偏差しか使われていない。

$$d = \frac{(\text{実験群の平均} - \text{統制群の平均})}{\sqrt{\frac{\text{実験群の標準偏差}^2 + \text{統制群の標準偏差}^2}{2}}} \qquad \text{式 (1)}$$

この計算から得られる値はグループごとの平均値の差を標準化したもの（standardized mean difference）になっている。算出される数値は，標準偏差を単位として平均値がどれだけ離れているかを表しており，たとえば，d = 1 なら，1 標準偏差（SD）分だけ離れていることを意味する。

表 1 のデータを例にしてみると，平均値差を 2 グループの標準偏差の平均で割れば 効果量 d が計算できる。単純化した式で書くと，3.00 ÷ [(10.00 + 10.00) ÷ 2] = 0.30 となる。つまり，グループ 1 の標準偏差（10.00）とグループ 2 の標準偏差（10.00）の平均が 10.00 であり，この標準偏差 1 つ分（1SD = 10.00）のうち，平均値差 3.00 の占める割合を見ているのである。表 1 では，効果量 d は両データとも 0.3 で効果量小（small effect size）という結果で，実質的な差は小さいということがわかる。

このように，効果量は，平均値と標準偏差のみでの直感的な判断とほとんど同じ解釈ができるものなのである。また，効果量は p 値のようにサンプル・サイズによって影響されることはないので，実質的な差を考えた場合には，統計的検定の枠組み（p 値）ではなく，効果量による解釈がふさわしいといえる。つまり，統計的検定の結果を解釈する際には，p 値を判断の最終材料とするべきではなく，まずは平均値，標準偏差，そして効果量によって，実質的な差を検討すべきである。また，研究における実験条件によっては，「有意差があっても（$p < .05$）効果量が小さい場合」もあれば，「有意差がなくても（$p > .05$）効果量が大きい場合」も考えられるため，有意差があろうがなかろうが，どちらにしても効果量は報告しなければならない（American Psychological Association, 2009; Field, 2009; Kline, 2004 など）。

よくある疑問としては，「効果量で実質的な差がわかるのであれば，統計的検定を行って p 値を見る必要はないのではないのか？」というものであるが，「効果量のみでよい」ということはない。そもそも，効果量は（母集団の特性を示そうする目的は同じであるが）確率を用いる推測統計とは目的が違うものであり，手元のデータから母集団にまで一般化を目指すのが統計的検定の目的なのである。データのサンプリングがうまくいっていないために，手元のデータが「たまたま」大きな差が得られるデータだったという場合は，効果量だけの解釈ではその可能性が否定できない。つまり，実質的な差を示す効果量が大きく，なおかつ統計的有意差もある（$p < .05$）というのが，理想的な統計的検定の形である。

2.2 効果量の指標と注意点

表2は検定・分析の種類別に代表的な効果量の指標と大きさの目安をまとめたものである（水

本・竹内，2008）。この表と効果量計算シート（http://www.mizumot.com/stats/effectsize.xls）は，発表されて以来，外国語教育学のみならず他分野でも利用されている。

水本・竹内（2008）では，「t 検定には繰り返しありと繰り返しなしのパターンがあるが，r と d ともに計算式は同じ形で効果量を求めることができる」（p. 63）という記述があるが，これについては間違った解釈を導いてしまう可能性があるため，説明を加える必要がある。なぜならば，繰り返しありの場合は，同一実験参加者が2度データ収集をされることになるため，データに対応が出てくる。そのため，効果量算出においてもデータの対応（相関係数）を考慮に入れるべきであるという考え方で計算されているものもある。以下の計算式 (2) は，後述の検定力分析を行うソフトである G*Power 3 で算出される，対応のある t 検定の場合の d である。式からわかるように，対応なしの場合の d を用いて，それを対応のあるデータの相関係数を用いることで調整している。

$$d_{Diff} = \frac{\text{対応なしの場合の} d}{\sqrt{2(1 - \text{対応のあるデータの相関係数})}} \qquad \text{式(2)}$$

一方，実験デザイン（対応のあり・なし）に関わらず，d は（対応のないときと）同じ値が得られるべきであるという考え方に基づいた計算方法もある。この計算方法がメタ分析で用いられていることからも（Borenstein, Hedges, Higgins & Rothstein, 2009），水本・竹内（2008）では，「r と d ともに計算式は同じ形で効果量を求めることができる」（p. 63）と記述した。しかし，r を用いた場合は，計算過程で繰り返しありの場合の t 値が使われており，データの対応（相関係数）を考慮に入れた値が計算されているため，対応のない場合の d とは違う意味を持った数値になっている。そのため，解釈には注意が必要になる。[4]

対応がある場合の d（もしくはそれに関連した指標）の計算式はいくつか存在するが，まとめると（a）対応のない場合の d と同じ値になるように計算しているもの（Cortina & Nouri, 2000, p. 49; Grissom & Kim, 2005, p. 67; Kline, 2004, p. 106）と，（b）データの対応を考慮して相関係数や平均値差で調整しているもの（Faul, Erdfelder, Lang & Buchner, 2007; Kline, 2004, p. 105, 豊田, 2009, p. 55）の2つに集約される。[5]

これらの d の値がデータ間の相関係数の変化によって，どのように変わるのかということを調べるために，次のようなシミュレーションを行った。まず，データ 1 (n = 100,000, Mean = 40, SD = 10) とデータ 2 (n = 100,000, Mean = 50, SD = 10) の 2 つのデータを，正規分布に従う形で 1,000 回発生させ，1 回ごとに効果量を計算し，最後に平均を出した（モンテカルロ・シミュレーション）。これらデータで，対応のない場合の d と同じ値が得られる計算式（グループ a）を用いると，必ず d = 1 になるはずので，2 つのデータの相関係数を 0.1 ずつ増加させていった場合に，対応のある d と同じ値が得られる計算式（グループ b）を使用すると，値がどのように変化するかを確認した（表 3 と図 1）。

表 2

検定・分析の種類別の代表的な効果量の指標と大きさの目安

使用される検定（分析）	対象と注意	効果量の指標	効果量の目安 小 (Small)	中 (Medium)	大 (Large)
相関分析		r	.10	.30	.50
重回帰分析		R^2	.02	.13	.26
		f^2	.02	.15	.35
t 検定（t-test）	r と d は対応ありの場合は注意	r	.10	.30	.50
		d	.20	.50	.80
一元配置分散分析 (One-way ANOVA)	全体の検定	η^2	.01	.06	.14
		partial η^2	-	-	-
		ω^2	.01	.09	.25
		f	.10	.25	.40
	多重比較	r	.10	.30	.50
		d	.20	.50	.80
二元配置分散分析 (Two-way ANOVA) 多元配置分散分析* (Multi-way ANOVA) *三元配置以上の分散分析	主効果	η^2	.01	.06	.14
		partial η^2	-	-	-
		ω^2	.01	.09	.25
	交互作用	η^2	.01	.06	.14
		partial η^2	-	-	-
		ω^2	.01	.09	.25
	多重比較	r	.10	.30	.50
		d	.20	.50	.80
共分散分析 (ANCOVA)	共変量の影響を取り除いて分析し，主効果，交互作用，多重比較の効果量は他の分散分析と同じ				
多変量分散分析 (MANOVA)	多変量検定	multivariate η^2 (multivariate R^2)	-	-	-
		multivariate partial η^2	-	-	-
多変量共分散分析 (MANCOVA)	従属変数ごとの分散分析	主効果，交互作用，多重比較の効果量は他の分散分析と同じ			
カイ 2 乗検定 (χ^2 test)	2×2 の分割表	φ (= w)	.10	.30	.50
	2×2 以外	Cramer's V	.10	.30	.50
＜ノンパラメトリック検定＞ マン・ホイットニーの U 検定 ウィルコクスンの符号順位和検定 クラスカル・ウォリスの順位和検定 フリードマン検定	検定統計量を Z に変換して r を求める	r	.10	.30	.50

水本・竹内（2008, p. 62）を基に作成。多重比較の場合，検定のように有意水準を調整する必要はない。

η^2 の大きさの目安は文献によっては，r を 2 乗した r^2 に合わせて，η^2 = .01（効果量小），η^2 = .09（効果量中），η^2 =.25（効果量大）としているものもある。また，partial η^2 の効果の大きさの基準は明確なものがない。multivariate η^2 と multivariate partial η^2 の値は従属変数（dependent variable）の数によって変わるため，効果量の目安は Cohen (1988) を参照。

表 3

2 つのシミュレーション・データの相関関係による効果量 *d* の変化

グループ	計算式	2 つのデータの相関係数								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(a)	対応のない場合の *d*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Borenstein et al. (2009, p. 29)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Cortina & Nouri (2000, p. 50)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(b)	式(2)で調整	0.75	0.79	0.85	0.91	1.00	1.12	1.29	1.58	2.24
	Kline (2004, p. 105)	0.75	0.79	0.85	0.91	1.00	1.12	1.29	1.58	2.24
	豊田 (2009, p. 55)	0.75	0.79	0.85	0.91	1.00	1.12	1.29	1.58	2.24

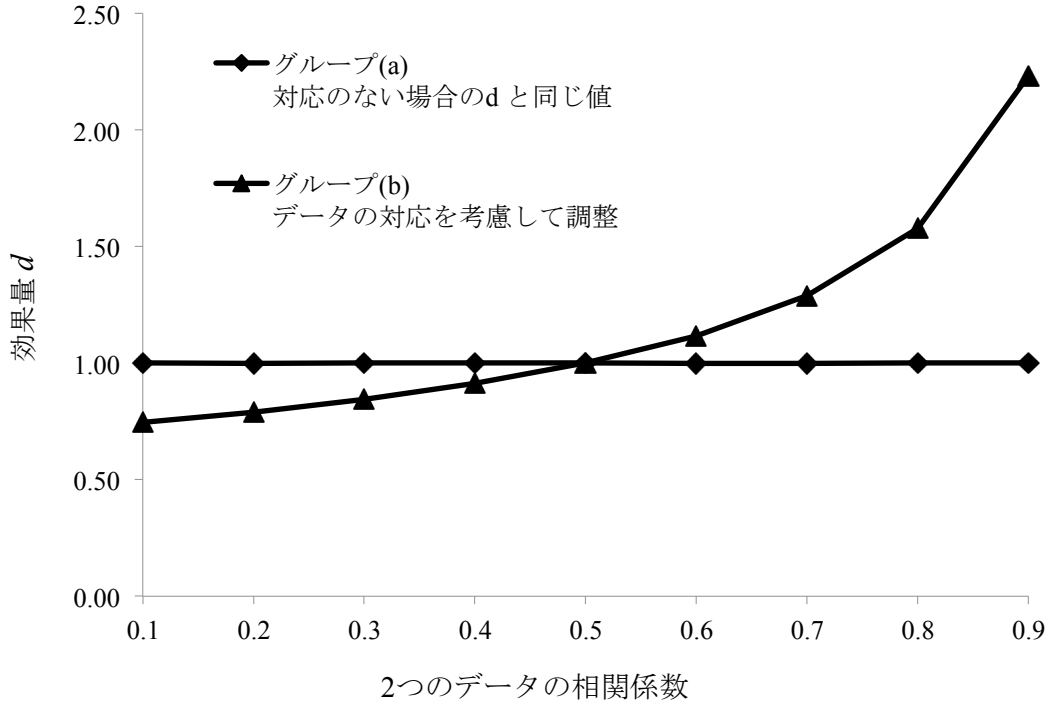


図 1　2 つのデータの相関関係による効果量 *d* の変化

図 1 からわかるように，2 つのデータの相関係数が変化しようとも，対応のない場合の d と同じ値が得られる計算式については，一定して $d = 1.00$ が得られている。一方，データの対応を考慮して相関係数や平均値差で調整している計算式は，2 つのデータの相関係数が 0.5 のときは，対応のない場合と同じ値になる（$d = 1$）。そして，0.5 より相関係数が小さい場合は，対応のない場合よりも値が小さくなり，相関係数が 0.5 より大きい場合は，対応のない場合よりも値が大きくなっている。2 つのデータの相関係数が 0.9 の場合は，対応のない場合の d 値（$d = 1$）に比べて 2 倍以上になっている（$d = 2.24$）。

これら 2 つの効果量の違いは，注目が「集団にあるのか（対応のない場合の d と同じ値）」，「個人の変化（データの対応を考慮して調整）にあるのか」の違いであり，どちらも有効な情報をもたらしていると考えればよい（豊田，2009，pp. 56–57）。

以上に示したシミュレーションの結果からわかるように，繰り返しのあるデータで効果量の d や r を報告するときは，どの計算式を使ったのかわかるように，参考文献や可能であれば式を明記しておくほうがよいだろう。また，繰り返しのあるデータの分析結果を論文で提示する際には，できる限り 2 つのデータの相関係数（もしくは差得点の平均値と標準偏差）を報告すべきである。相関係数が提示されていない場合は，後述する検定力分析でも，効果量を使って先行研究の結果を考察するメタ分析でも，データの対応を考慮して調整するグループ (b) の計算が正しく行えない（Dunlap, Cortina, Vaslow & Burke, 1996）。分析の再現性は量的研究の最低必要条件であるため，後から誰がデータ分析をしたとしても再現できるような結果の提示を心がけるべきである。

3. 検定力分析

3.1 統計的検定における 2 つの誤りと検定力

統計的検定においては，（a）サンプル・サイズ，（b）有意水準，（c）効果量，（b）検定力の 4 つが，検定結果の良し悪しを決定する要素である。これらの要素に関連して，統計的検定における「2 種類の誤り」について理解しておく必要がある。

有意水準は， 2.1 節で説明したとおり，実験の前に慣例として 5%に定めておき，実験で得られたデータから計算される p 値がその基準よりも小さければ，「有意差がある」と判断する。5%で設定されている有意水準は α（アルファ）で表される。有意水準は α を 5%と設定するということは，同時に，100 回中 5 回までは，推定を誤る可能性を認めている。つまり，$p < .05$ だからといって，常に有意差があるというわけではなく，「本当は有意差がないのに有意差がある」という誤った結論を下してしまう可能性を排除していないのである。そのため，有意水準 α は「実際には差がないのに差がある」と判断してしまう第 1 種の誤り（Type I error）を犯す確率を表している。

第 1 種の誤りのイメージを分かりやすく説明するために，コンセプトのみを捉えた

形で以下に図を使って説明する（あくまでコンセプトであるため，厳密な統計学の考え方に基づいたものではないことに注意していただきたい）。図 2 は検定を行ったあとに統計量として得られた値を「本当は 2 つのグループの平均値の差はないのが真実」というものさしを使って，どこに位置するかを測っている様子を表している（数値は実際には確率を表している）。グループの間の距離が離れているほど，平均値の差が大きいとして，実際に 2 つのグループに「差がない」場合には，(1) のようになる。

このものさしの右端に位置すればするほど，「差がないのが真実」のものさしでは測りきれないくらいの差があると考える（実際は，この差を作り出すのにサンプル・サイズが関係するため，実質的な差の大きさとイコールにはならないことに注意する。2.1 節参照）。どれほど右に行けば，「差がある」と考え始めることが可能かという基準が有意水準（α）であり，95%のところに基準を作る。すると残りは 5%になるので，ここに入るぐらいの差が得られた (2) のような場合は，「$p < .05$ で有意差がある」と解釈するとこの例では考えるようにする。

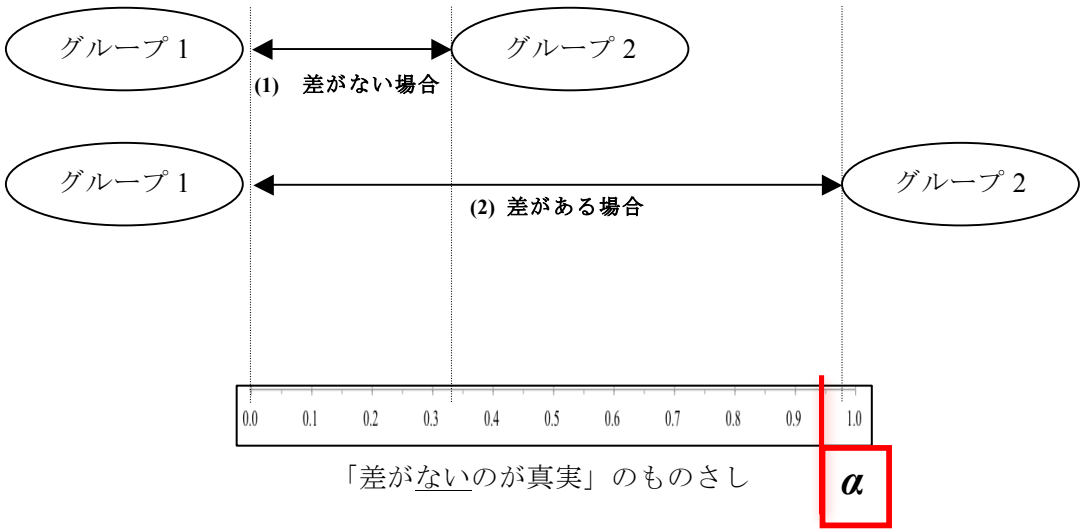


図 2　第 1 種の誤り（α）のイメージ

しかし，実際には差がないのに，たまたま設定した α よりも右の位置にくる値が得られることが 100 回に 5 回はある。その場合には，「差がないのに差がある」と誤った判断していることから，第 1 種の誤りを犯していることになる。

その一方で，「実際には有意差があるのに有意差なし」であるとしてしまう第 2 種の誤り（Type II error）も存在する。第 2 種の誤りの確率は β（ベータ）で表される。α は 0.05 と通常設定されているが，第 2 種の誤りを犯す確率は，$\beta = 0.20$（20%）が望ましい

とされている（Cohen, 1988）。「本当は有意差がないのに有意差がある」といってしまう第 1 種の誤りに比べると，「本当は有意差があるのに有意差がない」と判断するのは罪が軽いと考えられるため，α のように 0.05 ではなく，β は 0.20 とゆるめに設定されると考えればよい。第 1 種の誤りは，「差がないのが真実」のものさしを使っていたときに問題になっていたが，第 2 種の誤りは，図 3 下のような「差があるのが真実」のものさしを使うと考える。

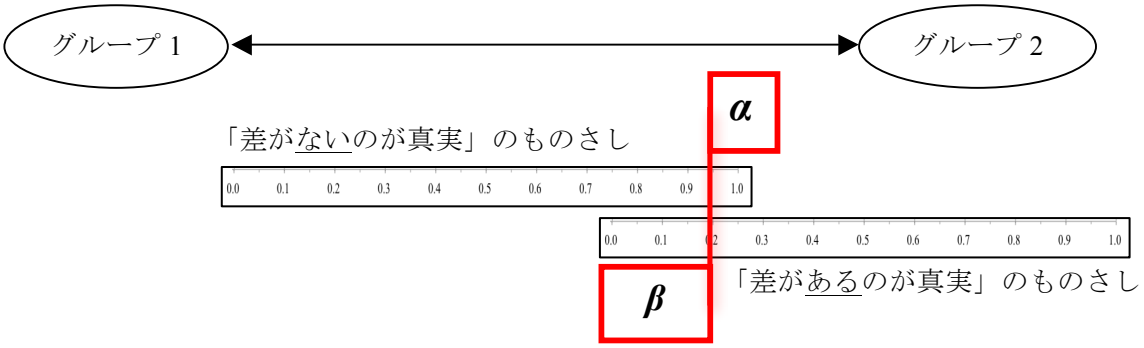


図 3　第 2 種の誤り（β）のイメージ

第 1 種の誤り（α）と第 2 種の誤り（β）は図 3 からもわかるように，同一のものさしを使っていないため，$\alpha + \beta = 1.0$ とはならない。しかし，拮抗する性質があるため（豊田, 2009, p. 31），第 1 種の誤りを犯す確率の α を小さく設定すると，β が大きくなってしまう。つまり，どちらか一方に注意を向ければ良いというものではなく，同時に考えなければならない（ただし，後述の検定力が高くなれば，β のみを小さくすることができる）。図 3 の中の α を右にずらした場合（有意水準を小さく設定した場合），β が大きくなることがわかるだろう。

「差があるのが母集団の本当の状態である」という前提で，実際に有意差を正しく検出できた場合には，統計的検定の目的が達成されているといえる。このように有意差を正しく検出できる確率のことを，「検定力」もしくは「検出力」（power）という。検定力は $1-\beta$ で定義される。つまり，「本当は差があるのに，差がない」と判断してしまう確率の β を 1 から引くことで，残りの「本当は差があり，差がある」と判断する確率を表している（図 4）。例えば，Cohen（1988）が推奨している β = 0.2 の場合，1−0.2 で 0.8 になる。検定力が 0.8 ということは，実際に有意差があるときには，80%の確率でそれを検出できることを意味している。また，Cohen（1992）は，「0.80 以下の検定力の場合には，第 2 種の誤りを犯す可能性が高くなる」（p. 156）としていて，検定力と第 2 種の誤りは表裏一体の関係にあることがわかる。

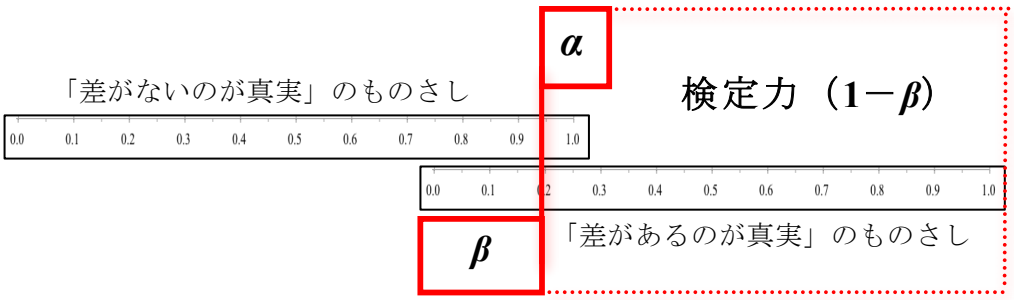


図 4　第 1 種の誤りと第 2 種の誤り，検定力の関係

3.2　検定力分析の目的

サンプル・サイズ，有意水準（α），検定力（1−β），効果量の 4 つは，他の 3 つが決まれば残りの 1 つが決まるという関係である。前節のように，推奨される検定力（1−β）は 0.8，有意水準（α）は 0.05 と決まっているので，[6] 統計的検定を用いる研究を計画する際に実際に考慮しなければならないのは，サンプル・サイズと効果量になってくる。

サンプル・サイズが，研究においてどれくらい必要かという疑問に答えることができるのが検定力分析である。サンプル・サイズが小さすぎると，検定力が下がってしまう可能性があり，大きすぎると検定力が大きくなりすぎて，実質的な差がなくても有意差があると判断してしまう。例えば，サンプル・サイズが非常に大きい場合などは，「手元の比較的小さなサンプルから，母集団の特性を推測する」という，統計的検定のそもそもの目的とは離れた行為になってしまう。つまり，できるだけ小さなサンプル・サイズで，検定力を大きく（第 2 種の誤りを小さく）し，検定を行うことが理想的といえる（豊田, 2009, p. 35）。

このような観点から，実験を行う前に検定力分析（power analysis）を利用し，サンプル・サイズを決定することが推奨される。検定力分析は，おもに以下の 2 つの目的で行われる（その他の検定力分析は，Faul, Erdfelder, Lang, & Buchner, 2007 を参照）。

(1) サンプル・サイズを決める（事前の分析: A priori）
　実験を実施する前に，これまでの先行研究からわかっている（推測される）効果量，有意水準（α），目指している検定力（1−β）からサンプル・サイズを決定する。

(2) 検定力を調べる（事後の分析: Post hoc）
　実験を実施した後に，サンプル・サイズ，効果量，有意水準（α）から，検定力（1−β）を確認する。

検定力分析は，図 5 の統計的検定における 4 つの要素では，3 つが決まれば残り 1 つが決まる関係にある性質を利用して求めることができる。例えば，上記の検定力分析の (1) で実験前にサンプル・サイズを決める場合は，図 5 のサンプル・サイズ以外の 3 つ（有意水準，検定力，効果量）の値を使えば計算することが可能である。また，(2) の検定力を事後の分析で求める場合も，図 5 の検定力以外の 3 つ（有意水準，サンプル・サイズ，効果量）の値を使えばよい。

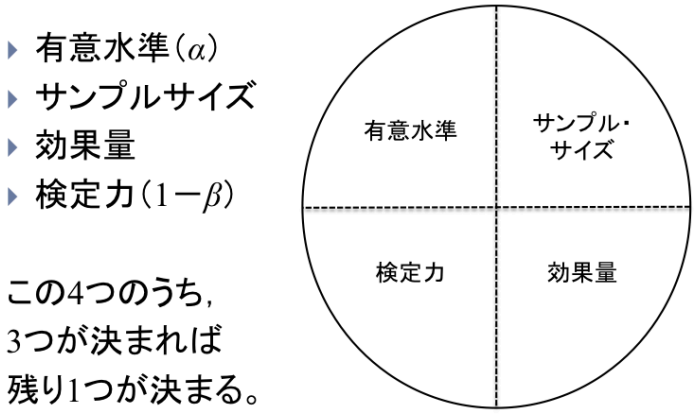


図 5　統計的検定における 4 つの要素

3.3　G*Power 3 を使った各種検定の検定力分析

検定力分析は，ネット上で入手できるフリーソフト G*Power 3（Faul, Erdfelder, Lang & Buchner, 2007: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/）[7] を使って比較的簡単に実行できる（図 6 のように G*Power 3 は Mac 版も Windows 版も提供されている）。以下では，G*Power 3 を使って，主要な各種検定の検定力分析（事前と事後）を行う方法を説明する（2011 年 3 月現在，G*Power 3 のホームページでは説明が未だに不十分であるため有用であると思われる）。なお，検定力を実験後に分析する事後の分析（Post hoc）では，得られる情報が少ないと考えられているため（Hoening & Heisey, 2001; O'Keefe, 2007），検定力分析の実際の適用例は，より重要な実験前のサンプル・サイズの計画である事前の分析（A priori）のみに限定した。

＜G*Power 3 での検定力分析の方法を本稿で説明する各種検定＞

(1) 対応なしの *t* 検定 (independent *t*-test)

(2) 対応ありの *t* 検定 (dependent *t*-test)

(3) 対応なしの一元配置分散分析 (one-way ANOVA)

(4) 対応ありの一元配置分散分析 (one-way repeated measures ANOVA)

(5) 二元配置分散分析 (two-way ANOVA)

(6) 共分散分析 (ANCOVA)

(7) 多変量分散分析 (MANOVA)

(8) カイ 2 乗検定 (χ^2 test)

(9) ノンパラメトリック検定 (nonparametric tests)

(10) 相関係数 (correlation)

(11) 単回帰・重回帰分析 (regression analysis)

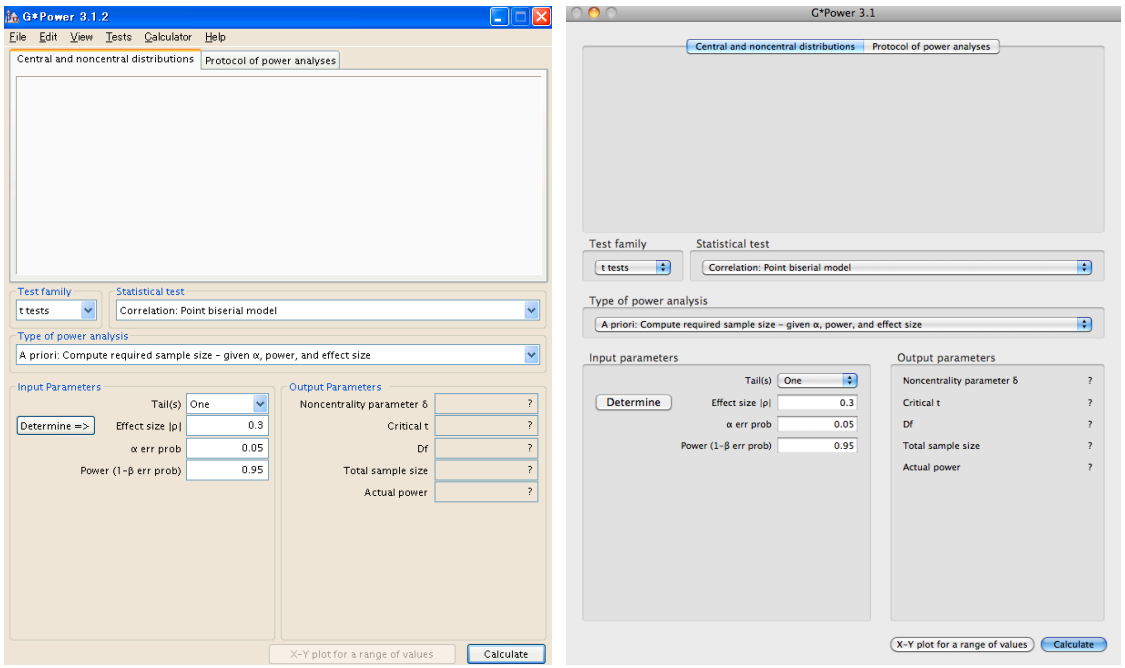


図 6　G*Power 3 起動時（左が Windows 版，右が Mac 版）

(1) 対応なしの *t* 検定（independent *t*-test）

Test family t tests

Statistical test Means: Difference between two independent means (two groups)

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Tails(s): Two（普通は両側検定）

Effect size d: 先行研究からわかっている効果量の大きさを入力

※もし，先行研究での効果量がわからなければ，d = 0.2（効果量小），0.5（効果量中），0.8（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.5（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8（Cohen, 1992 で推奨されている検定力）

Allocation ratio N2/N1: 2 つのグループの n の比

・事後の分析（Post hoc）の場合

Tails(s): Two（普通は両側検定）

Effect size d: 得られたデータの効果量。“Determine”をクリックすると，Mean や SD を入力することで効果量を計算できる。

※この d は本稿表 4 のグループ B の計算で求められたもの。

α error prob: 有意水準 0.05

Sample size group 1: グループ 1 の人数（サンプル・サイズ）

Sample size group 2: グループ 2 の人数（サンプル・サイズ）

適用例　事前の分析（A priori）

両側検定，中程度の効果量（d = 0.5），α = 0.05，Power = 0.8，Allocation ratio = 1

→以上の条件で，サンプル・サイズは各群 64 名（計 128 名）必要。

(2) 対応ありの *t* 検定（dependent *t*-test）

Test family t tests

Statistical test Means: Difference between two dependent means (matched pairs)

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Tails(s): Two（普通は両側検定）

Effect size d: 先行研究からわかっている効果量の大きさを入力

※もし，先行研究での効果量がわからなければ，d = 0.2（効果量小），0.5（効果量中），0.8（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.5（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 推奨されている検定力 0.8（Cohen, 1992）

・事後の分析（Post hoc）の場合

Tails(s): Two（普通は両側検定）

Effect size d: 得られたデータの効果量。この d は表 4 のグループ B の計算で求められたもの。“Determine”をクリックすると，Mean や SD を入力することで効果量を計算できる（グループ間の相関係数が必要）。

※この d は表 4 のグループ B の計算で求められたもの。

α error prob: 有意水準 0.05

Total sample size: 実験参加者数（サンプル・サイズ）

適用例　事前の分析（A priori）

両側検定，中程度の効果量（d = 0.5），α = 0.05，Power = 0.8

→以上の条件で，サンプル・サイズは 34 名必要（対応なしの場合の約半分になっている）。

(3) 対応なしの一元配置分散分析（one-way ANOVA）

Test family F tests

Statistical test ANOVA: Fixed effects, omnibus, one-way

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Effect size f: 先行研究からわかっている効果量の大きさを入力する。もし，先行研究での効果量がわからなければ，f = 0.10（効果量小），0.25（効果量中），0.40（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.25（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8

Number of groups: グループの数

・事後の分析（Post hoc）の場合

Effect size f: 得られたデータの効果量を計算する。“Determine”をクリックすると，平均や分散，partial η^2 などから効果量 f を計算できる。

α error prob: 有意水準 0.05

Total sample size: すべてのグループの人数の合計

Number of groups: グループの数

適用例　事前の分析（A priori）

中程度の効果量（f = 0.25），α = 0.05，Power = 0.8，3 群

→以上の条件で，サンプル・サイズは合計 159 名必要（1 群あたり 53 名[159 名/3 群]）。

(4) 対応ありの一元配置分散分析（one-way repeated measures ANOVA）

Test family F tests

Statistical test ANOVA: Repeated measures, within factors

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Effect size f: 先行研究からわかっている効果量の大きさを入力する。もし，先行研究での効果量がわからなければ，f = 0.10（効果量小），0.25（効果量中），0.40（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果

量を入力しておく。もし何もわからなければ，0.25（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8

Number of groups: グループの数（一元配置なので 1 を入力）

Number of measurements: 水準の数（繰り返した測定の数）

※例えば，pre, post, delayed と 3 回測定を行った場合は，3 と入力する。

Corr among rep measures: 水準間の相関

※先行研究やパイロットスタディで相関がわかるのであればそれを入力。もしわからなければ，0.5 としておく。

Nonsphericition correction ε: 球面性の仮定（sphericity assumption）が満たされていたら 1

※球面性の仮定が満たされていない場合は，1/(水準数−1) で下限値を入力しておけばよい（Faul, Erdfelder, Lang & Buchner, 2007, p. 181）。MANOVA で検定力を求める方法の場合は，球面性の仮定は必要ない（→MANOVA: Repeated measures, within factors で実行できる）。

・事後の分析（Post hoc）の場合

Effect size f: 得られたデータの効果量を計算する。“Determine”をクリックすると，partial η^2 などから効果量 f を計算できる。

α error prob: 有意水準 0.05

Total sample size: すべてのグループの人数の合計

Number of groups: グループの数（一元配置の場合は 1）

Number of measurements: 水準の数（繰り返した測定の数）

※例えば，pre, post, delayed と 3 回測定を行った場合は，3 と入力する。

Corr among rep measures: 水準間の相関（級内相関）

Nonsphericition correction ε: 球面性の仮定（sphericity assumption）が満たされていたら 1

※球面性の仮定が満たされていない場合は，1 / (水準数-1) で下限値を入力しておけばよい（Faul, Erdfelder, Lang & Buchner, 2007, p. 181）。SPSS などでは，「イプシロン」として Greenhouse-Gaisser（グリーンハウス・ゲイザー）や，Huynh-Feldt（ホイン・フェルト）の値が出力されるので，そちらの数値を使えば正確な値が得られる。

適用例　事前の分析（A priori）

中程度の効果量（f = 0.25），α = 0.05，Power = 0.8，1 群（一元配置のため），3 回の測定（pre, post, delayed など），相関は 0.5，Nonsphericition correction ε = 1

→以上の条件で，サンプル・サイズ 28 名。

(5) 二元配置分散分析（two-way ANOVA）

二元配置以上（多要因）の分散分析では，要因の主効果と交互作用，それぞれに対して検定力分析を行うことになる。

Test family F tests

Statistical test （2 要因とも対応なしの場合）ANOVA: Fixed effects, special, main effects and interactions

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Effect size f: 先行研究からわかっている効果量の大きさを入力する。もし，先行研究での効果量がわからなければ，f = 0.10（効果量小），0.25（効果量中），0.40（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.25（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8

Numerator df: 水準数−1

※二元配置では 2 つの要因のうち水準の多い方を用いれば，より多い必要人数がわかる。交互作用は，(要因 A-1)×(要因 B-1) になる。主効果か交互作用のどちらを使ってサンプル・サイズを計算するかは，研究の目的による（交互作用があることが期待されて，それが研究の目的に関連している場合は，交互作用でのサンプル・サイズを算出する）。

Number of groups: 「グループの数」ではなく，要因 A の水準×要因 B の水準で計算される数（総セル数）。

・事後の分析（Post hoc）の場合

※要因の主効果と交互作用，それぞれに対して検定力分析を行う。

Effect size f: 得られたデータの効果量を計算する。“Determine”をクリックすると，partial η^2 などから効果量 f を計算できる。

α error prob: 有意水準 0.05

Total sample size: すべてのグループの人数の合計

Numerator df: 水準数−1（主効果，交互作用ごとに計算する）。交互作用は，(要因 A-1)×(要因 B-1) になる。

Number of groups:「グループの数」ではなく，要因 A の水準×要因 B の水準で計算される数（総セル数）。

適用例　事前の分析（A priori）

要因 A（3 水準）×要因（3 水準）の二元配置分散分析を行う場合。中程度の効果量（f = 0.25），α = 0.05，Power = 0.8，Numerator df = 2（両要因とも 3 水準なので 3-1），Number of groups = 9（3 水準×3 水準）

→以上の条件で，セル 1 つにつき 18 名必要（=合計 158 名/[3 群×3 水準]）。

＜追加説明＞

対応のない要因（3 水準）と対応のある要因（3 水準）の二元配置分散分析の場合

対応のない要因　ANOVA: Repeated measures, 2between factors で検定力算出
Effect size f: 0.25（効果量中）
α error prob: 0.05
Power (1–β error prob): 0.8
Number of groups: 3
Number of measurements: 3
Corr among rep measures: 0.5
→Total sample size: 108（1 群につき 36 名必要）

対応のある要因　ANOVA: Repeated measures, within factors で検定力算出
Effect size f: 0.25（効果量中）
α error prob: 0.05
Power (1–β error prob): 0.8
Number of groups: 3
Number of measurements: 3
Corr among rep measures: 0.5
Nonsphericition correction ε: 1
→Total sample size: 30（1 群につき 10 名必要）

交互作用　ANOVA: Repeated measures, within-between interaction で検定力算出
Effect size f: 0.25（効果量中）
α error prob: 0.05
Power (1–β error prob): 0.8
Number of groups: 3
Number of measurements: 3
Corr among rep measures: 0.5
Nonsphericition correction ε: 1
→Total sample size: 36（1 群につき 12 名必要）

(6) 共分散分析 (ANCOVA)

Test family F tests

Statistical test ANCOVA: Fixed effects, main effects and interactions

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Effect size f: 先行研究からわかっている効果量の大きさを入力する。もし，先行研究での効果量がわからなければ，f = 0.10（効果量小），0.25（効果量中），0.40（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.25（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8

Numerator df: グループの数−1

Number of groups: グループの数

Number of covariates: 使用する共変量の数

・事後の分析（Post hoc）の場合

Effect size f: 得られたデータの効果量を計算する。“Determine”をクリックすると，partial η^2 などから効果量 f を計算できる。

α error prob: 有意水準 0.05

Total sample size: すべてのグループの人数の合計

Numerator df: グループの数−1

Number of groups: グループの数

Number of covariates: 使用する共変量の数

適用例　事前の分析（A priori）

中程度の効果量（f = 0.25），α = 0.05，Power = 0.8，3 群，共変量 1 つ

→以上の条件で，サンプル・サイズは合計 158 名必要（1 群あたり 53 名×3 群）。

(7) 多変量分散分析 (MANOVA)

Test family F tests

Statistical test MANOVA: Global effects（一元配置多変量分散分析モデルの場合）

※ 二元配置以上の多変量分散分析の場合は，目的に応じて以下を使用（二元配置分散分析の説明を参照）。

MANOVA: Special effects and interactions

MANOVA: Repeated measures, between factors

MANOVA: Repeated measures, within factors

MANOVA: Repeated measures, within-between interaction

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Effect size f^2(V): 先行研究からわかっている効果量の大きさを入力する。もし，先行研究での効果量がわからなければ，f^2= 0.02（効果量小），0.15（効果量中），0.35（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.15（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8

Number of groups: グループの数

Response variables: 従属変数の数

・事後の分析（Post hoc）の場合

Effect size f^2(V): 得られたデータの効果量を計算する。“Determine”をクリックすると，Pillai V（Options でその他の効果量を選ぶことも可能）から効果量 f^2 を計算できる。

α error prob: 有意水準 0.05

Total sample size: 人数の合計

Number of groups: グループの数

Response variables: 従属変数の数

適用例　事前の分析（A priori）

中程度の効果量（$f^2 = 0.15$），$\alpha = 0.05$，Power = 0.8，3 群，従属変数 2 つ

→以上の条件で，サンプル・サイズは合計 45 名必要（1 群あたり 15 名×3 群）。

※ MANOVA であれば，上記のように 1 群あたりの人数は比較的少なくて構わないが，同じサンプル・サイズで MANOVA のあとに ANOVA を行うような場合には，MANOVA の基準では検定力が非常に低くなってしまうため，実験計画段階で後の ANOVA までを考慮してサンプル・サイズを算出しておかなければならない。

(8) カイ 2 乗検定（χ^2 test）

Test family χ^2 tests

Statistical test Goodness-of-fit tests: Contingency tables

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Effect size w: 先行研究からわかっている効果量の大きさを入力

※もし，先行研究での効果量がわからなければ，w = 0.1（効果量小），0.3（効果量中），0.5（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.3（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8

Df: 自由度　・1 変数を扱う適合度検定の場合は「カテゴリ数−1」となる。

・クロス表の検定である独立性検定では「(行の数−1)×(列の数−1)」となる。

・事後の分析（Post hoc）の場合

Effect size w: 得られたデータの効果量。

α error prob: 有意水準 0.05

Total sample size: 合計人数

Df: 自由度　・1 変数を扱う適合度検定の場合は「カテゴリ数−1」となる。

・クロス表の検定である独立性検定では「(行の数−1)×(列の数−1)」となる。

適用例　事前の分析（A priori）

中程度の効果量（w = 0.3），α = 0.05，Power = 0.8，3×4 のクロス表

→以上の条件で，サンプル・サイズは 152 名必要。

(9) ノンパラメトリック検定 (nonparametric tests)

G*Power 3 では，ノンパラメトリック検定の検定力は，マン・ホイットニーの U 検定（2 群，データの対応なし）とウィルコクスンの符号順位和検定（2 群，データの対応あり）が用意されているが，その他の場合は，対応するパラメトリック検定で検定力分析を行えばよい。ただし，パラメトリック検定の前提を満たしている場合に，ノンパラメトリック検定を行うと検定力が下がる（Siegel & Castellan, 1988）。

＜2 群の場合＞

マン・ホイットニーの U 検定（2 群，データの対応なし）

t tests → Means: Wilcoxon-Mann-Whitney test (two groups)

ウィルコクスンの符号順位和検定（2 群，データの対応あり）

t tests → Means: Wilcoxon signed-rank test (matched pairs)

＜3 群以上の場合＞

クラスカル・ウォリスの順位和検定（3 群以上，データの対応なし）

F tests → ANOVA: Fixed effects, omnibus, one-way

フリードマン検定（3 群以上，データの対応あり）

F tests → ANOVA: Repeated measures, within factors

(10) 相関係数 (correlation)

相関係数の検定は「無相関検定」と呼ばれており，「母集団の相関が 0 である（$\rho = 0$）」という帰無仮説を検定する。相関係数の強さには関係がないことからも，この検定自体にはほとんど意味がない（前田, 2004, p. 66）。つまり，「母相関が 0 ではないか」という点を，手元のデータの相関係数から行えるだけのサンプル・サイズがあるかということだけを確認できる。

Test family Exact

Statistical test Correlation: Bivariate normal model

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Tails(s): Two

Correlation ρ H1: 先行研究からわかっている効果量の大きさを入力

※もし，先行研究での効果量がわからなければ，ρ = 0.1（効果量小），0.3（効果量中），0.5（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.3（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8

Correlation ρ H0: 0（検定の帰無仮説「母相関は 0 である」という値を入力）

・事後の分析（Post hoc）の場合

Tails(s): Two

Correlation ρ H1: 得られたデータの効果量（相関係数の場合は，r をそのまま入力）

α error prob: 有意水準 0.05

Total sample size: サンプル・サイズ

Correlation ρ H0: 0（検定の帰無仮説「母相関は 0 である」という値を入力）

適用例　事前の分析（A priori）

両側検定，中程度の効果量（$\rho = 0.3$），$\alpha = 0.05$，Power = 0.8

→以上の条件で，サンプル・サイズは 84 名必要。

(11) 単回帰・重回帰分析 (regression analysis)

回帰分析では，回帰係数の検定を行うが，この検定は（10）で説明している相関係数の無相関検定と同じく，「母集団での回帰係数（β）が 0 である」という帰無仮説を検定する。以下は単純な回帰分析の場合の G*Power 3 での検定力分析であるが，その他の回帰分析の場合は，Faul, Erdfelder, Buchner and Lang (2009) を参照。

Test family F tests

Statistical test Linear multiple regression: Fixed model, R^2 deviation from zero

※単回帰・重回帰分析ともにできるが，重回帰の決定係数の増分に関する検定については，Linear multiple regression: Fixed model, R^2 increase を使用する。Number of tested predictors に増加分の説明変数の数を入れて，Total

number of predictors に全説明変数の数を入れれば分析可能。

Type of power analysis

・事前の分析の場合　A priori: Compute required sample size – given α, power, and effect size

・事後の分析の場合　Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

・事前の分析（A priori）の場合

Effect size f^2: 先行研究からわかっている効果量の大きさを入力する。もし，先行研究での効果量がわからなければ，$f^2 = 0.02$（効果量小），0.15（効果量中），0.35（効果量大）の Cohen（1988）の基準を用いて，自分の研究での予測される効果量を入力しておく。もし何もわからなければ，0.15（効果量中）にしておく。

α error prob: 有意水準 0.05

Power (1–β error prob): 0.8

Number of predictors: 説明変数の数（単回帰の場合は 1，重回帰の場合は 2 以上）

・事後の分析（Post hoc）の場合

Effect size f^2: 得られたデータの効果量

α error prob: 有意水準 0.05

Total sample size: サンプル・サイズ

Number of predictors: 説明変数の数（単回帰の場合は 1，重回帰の場合は 2 以上）

適用例　事前の分析（A priori）

中程度の効果量（$f^2 = 0.15$），$\alpha = 0.05$，Power = 0.8，説明変数 3 つ

→以上の条件で，サンプル・サイズは 77 名必要。

4. まとめ

本稿では，効果量がなぜ必要なのかを説明し，いくつかある効果量の指標の中で，対応のあるデータにおける *d* の算出方法について，シミュレーションを用いながら検証を行った。次に，検定力分析の基礎的な考え方と，G*Power 3 を使った検定力分析の具体的な方法の解説を行った。

効果量と検定力分析という，統計的検定における重要概念は，いくつかの応用言語学関連の国際ジャーナルでも論文中で報告するようにと推奨されているため，今後はさらに利用が増えていくはずである。たとえば，我々の分野における有力国際誌の *TESOL Quarterly* における Quantitative Research Guidelines（量的研究ガイドライン）では次のよう

なセクションがある（http://www.tesol.org/s_tesol/sec_document.asp?CID=476&DID=1032）。

> *Power and sample size.*
> Provide information on the sample size and the process that led to the decision to use that size. Provide information on the anticipated effect size as you have estimated it from previous research. Provide the alpha level used in the study, discussing the risk of Type I error. Provide the power of your study (calculate it using a standard reference such as Cohen, 1988, or a computer program). Discuss the risk of Type II error.

このガイドラインからも，本稿で説明した，効果量や検定力（そして，第1種の誤りと第 2 種の誤り）を理解し，活用していくことは非常に重要であるといえるだろう。

最後に，効果量と検定力分析の概念がわかっていれば，論文中で提示されているデータへの洞察力が深まるということを示す例を一つ挙げておく。以下の表 4 は，コンピュータを使った学習・指導についての研究を専門にしている，ある有名な国際ジャーナルに掲載されていた論文から抜粋したものである（表の見た目は手を加えてあるが，p 値の記載方法も含めて，数値はそのままにしている）。この「統計的に有意な差が見られた」という結果から，著者はコンピュータを使った学習（処置群）が，使わなかった学習（対照群）よりも効果的であったという主張を行っている。

表 4
ある研究論文に掲載されていた結果

グループ	人数	テストの点数（平均点）	標準偏差	対応のない t 検定
処置群（treatment）	11	78.91	8.42	$p = 0.000$
対照群（contrast）	11	76.82	7.41	

この結果を効果量で解釈すると，$d = 0.26$（効果量小）となり，$p = 0.000$ という p 値にも関わらず，効果量は小さいということがわかる。また，検定力の事後の分析（Post hoc）では，Power（$1-\beta$）が 0.09 となり，非常に検定力が低い検定であったこともわかる。同じ効果量で，$\alpha = .05$，検定力 0.8 を得るためには，サンプル・サイズは各グループに 228 名必要であることからも，そもそも $p = 0.000$ という値は得られないのではないかという疑問が湧いてくる。そこで，同じ人数，平均値，標準偏差のデータを再現し，シミュレーションを行ってみたところ $p = .54$ であった。つまり，論文で報告されている p 値は間違い（もしくは偽り）なのである。山森（2004）が述べているように，「査読者も編集者も万能ではない」（p. 158）というのが，このような結果が堂々と国際ジャーナルに掲載されている原因になっていると思われるが，このような論文が 1 本掲載されているだけで，ジャー

ナル自体の価値が下がってしまうといっても過言ではないため，査読者や編集者の責任は重いだろう。

このように，効果量と検定力分析が使用できれば，p 値のみで間違った結果の解釈を行っている論文に警笛を鳴らすことも可能なのである。外国語教育学研究でも，効果量と検定力分析の考え方が広まり，論文での報告が増え，「正しい統計的検定」が行われることを期待している。

謝辞

本稿は，平成 20 年度～22 年度科学研究費補助金（基盤研究(C)「外国語学習方略の脳内基盤：読解方略の意識化と指導モデルの視点から」課題番号：20520540，研究代表者：関西大学外国語学部 竹内 理）の内容の一部を基にしたものである。また，内容については，印南 洋氏（豊橋技術科学大学），小泉利恵 氏（常磐大学）から貴重なアドバイスを頂いた。ここに記して感謝する。

注

1. 有意水準は検定前に一定の値に決めておく基準で，p 値は検定後に得られる具体的な数値であって 2 つは似ているが同じものではない。また，有意水準を .05 に設定するというのは全く恣意的（そして慣例的）なものであるので，絶対的な基準ではない。
2. この乱数は，群馬大学の青木繁伸先生のホームページ（http://aoki2.si.gunma-u.ac.jp/R/misc2.html）を参考にして，R で発生させた。
3. Kline (2004, p. 102) は，この計算によって得られる値は Hedges's g であり，厳密には，この指標を d と呼ぶのは間違いであるとしている。しかし，d として使われることが圧倒的に多いため，本稿でもこの指標を d として扱う。
4. 繰り返しのある場合の効果量 r は，繰り返しのある場合の t 検定の t 値を計算式に用いるが，データの相関により，実際の効果量よりは高めの値が算出されることになる（Field, 2009, p. 332）。また，一般に知られている，効果量 r から d への変換式である，$d = 2r / \mathrm{sqrt}\ (1 - r^2)$ は，対応のない場合の計算式であるので，対応のある場合に正確な d には変換できない。
5. 比較した計算式は以下のもの。

Borenstein, et al. (2009, p. 29) $$d = \frac{\text{平均値差}}{\left(\frac{\text{平均値差の}SD}{\sqrt{2(1-r)}}\right)}$$

Cortina & Nouri (2000, p. 50) $$d = |\text{対応ありの}t\text{値}| \times \sqrt{\frac{2(1-r)}{n}}$$

Kline (2004, p. 107) $$d = \left|\text{対応ありの}t\text{値}\right| \times \sqrt{\frac{2 \times \text{平均値差の}SD}{n(\text{データ 1 の}SD^2 + \text{データ 1 の}SD^2)}}$$

Grissom & Kim (2005, p. 67); Kline (2004, p. 105) $$d = \frac{\text{平均値差}}{\text{平均値差の}SD}$$

豊田 (2009, p. 55) $$d = \left|\text{対応ありの}t\text{値}\right| \times \sqrt{\frac{1}{n}}$$

6. これらは推奨されている値であるため，もちろん，研究の目的によっては違う値に変えても構わない。
7. 検定力分析は R でも実行可能（豊田, 2009; 山田・杉澤・村井, 2008）。

参考文献

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* West Sussex, U.K. John Wiley & Sons.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage.

Dunlap, P. W., Cortina, M. J., Vaslow, B. J., & Burke, J. B. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. Retrieved from http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power : The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24. Retrieved from www.vims.edu/people/hoenig_jm/pubs/hoenig2.pdf

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

前田啓朗 (2004).「テスト得点間の関係の検討―相関分析―」. 前田啓朗・山森光陽 (編)磯田貴道・廣森友人(著)『英語教師のための教育データ分析入門:授業が変わるテスト・評価・研究』(pp. 64–72). 東京:大修館書店.

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. Retrieved from http://www.mizumot.com/files/EffectSize_KELES31.pdf

O'Keefe, D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, *1,* 291–299. Retrieved from http://www.dokeefe.net/pub/OKeefe07CMM-posthoc.pdf

Siegel, S., & Castellan, N. J. Jr. (1988). *Nonparametric statistics for the behavioral sciences.* (2nd ed.). New York: McGraw-Hill.

豊田秀樹(編著)(2009).『検定力分析入門―R で学ぶ最新データ解析―』東京:東京図書.

山田剛史・杉澤武俊・村井潤一郎 (2008).『R によるやさしい統計学』東京:オーム社.

山森光陽 (2004).「分析結果の書き方ガイド」. 前田啓朗・山森光陽(編)磯田貴道・廣森友人(著)『英語教師のための教育データ分析入門:授業が変わるテスト・評価・研究』(pp. 158–174). 東京:大修館書店.