統計数理研究所共同研究リポート 238
『言語コーパス分析における数理データの統計的処理手法の検討』(2010) pp. 1–14

サンプルサイズが小さい場合の統計的検定の比較
―コーパス言語学・外国語教育学への適用―

水本　篤
流通科学大学
E-mail:　atsushi@mizumot.com

あらまし　本研究では，サンプルサイズが小さい場合に使用される統計的検定の比較を行った。特に，従来から使用されることが多いパラメトリック検定とノンパラメトリック検定，そして，並べ替え検定とフィッシャーの正確確率検定という正確な p 値を得ることができる方法を比較することを目的とした。

キーワード　並べ替え検定，確率化検定，フィッシャーの正確確率検定，検定力（分析）

A Comparison of Statistical Tests for a Small Sample Size: Application to Corpus Linguistics and Foreign Language Education and Research

Atsushi MIZUMOTO
University of Marketing and Distribution Sciences

Abstract　This paper reports on a comparison of statistical tests for a small sample size. First. I will explain the characteristics of conventional parametric and nonparametric procedures. In addition to these two approaches, comparisons will be made on permutation tests and Fisher's exact tests with examples.

Keyword　permutation test, randomization test, Fisher's exact test

1. はじめに

コーパス言語学で扱う言語データはサンプルサイズ（n）の数が大きいことが多いので，統計的検定の仕組みから（詳細は後述を参照），検定するまでもなく「差がある」という結果になるであろうということが多い。統計的検定（推測）は，そもそも母集団すべてのデータを収集することは不可能であるから，少ないサンプルサイズで，その母集団の分布を推定することを目的としているため，サンプルサイズが小さい場合こそ，利用の価値が高い方法であるといえるだろう。

コーパス言語学だけではなく，外国語教育学などの分野においても，実験・研究によって少ないサンプルサイズしか得られないことがある。そのような場合にふさわしい検定はどのようなものがあり，どのような特徴があるのか，本稿では例を示しながら検討していく。

2. 統計的検定における 4 つの重要な要素

まず，本稿の目的である，いくつかの検定の比較を行う前に，統計的検定における 4 つの重要な要素を説明しておく。「4 つの重要な要素」とは，有意水準（α），検定力（$1-\beta$），サンプルサイズ，効果量のことである。Field and Hole（2003）や 村井（2006）によると，有意水準（α），検定力（$1-\beta$），サンプルサイズ，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とされている。

2.1 有意水準（α）

統計的検定では「統計的に有意な差がある」ということを示すために行うため，はじめに「差がない」という帰無仮説（null hypothesis）を設定する。そして，慣例として，検定結果が $p < .05$（5 %以下）[1]であった場合には，100 回中 5 回以下の低い確率で「差がない」という事象が偶然起こることを表すので，「差がないといえる確率はほとんどない」と考える。そのため，設定しておいた帰無仮説を棄却し「差がある」という対立仮説（alternative hypothesis）を採択し，「統計的に有意な差がある」という判断を下す。

ここで注意しなければならないのは，$p < .05$ で有意な差があると判断しても，その判断が誤りである確率が p 値と同じだけあるということである。例えば，$p = .05$ であったとしても，母集団の性質の推定を行ったときに，20 回に 1 回は判断が外れる可能性があることを認めているのが統計的有意差検定の考え方である。このような理由から，α で表される有意水準は「実際には有意差がないのに有意差がある」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おり，「危険率」とも呼ばれる。

---

[1] 表記は 5%「未満」を表す $p < .05$ となっているが，$p = .05$ の場合も厳密には「5%水準で有意差あり」という基準に含まれるため，「5%以下」という表現を用いる。なお，数値が 1%を超えない場合は .05（=0.05）のように小数点前の 0 を書かないのが慣例であるため（APA, 2009），本稿でもそのように表記する。

2.2 検定力（Power）

第 1 種の誤りに加えて，「実際には有意差があるのに有意差なし」としてしまう第 2 種の誤り（Type II error）を犯す確率（β）の可能性も常に考えなければならない。有意水準や危険率とも呼ばれる第 1 種の誤り（Type I error）を犯す確率（α）は，（心理学などの分野で）統計的検定の慣例として，$\alpha = .05$ に決められており，第 2 種の誤り（Type II error）を犯す確率（β）は，研究者自身が設定することが目的に応じて設定できるが，$\beta = .20$ が望ましいと提案されていている（Cohen, 1988）。そして検定力（Power）は $1-\beta$ で定義されるため，$\beta = 0.2$ の場合，$1-0.2$ で検定力が 0.8 になる。検定力が 0.8 ということは，80%の確率で実際に有意差があるときには，それを正しく検出できることを意味している。また，Cohen（1992, p. 156）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」としている。このように，第 2 種の誤りは検定力の計算に直接関わってくる問題である。

表 1　統計的有意差検定における 2 種類の誤りと検定力

検定結果 / 真実	差がないと判断 （帰無仮説を採択）	差があると判断 （帰無仮説を棄却）
本当は差がない	正しい判断	第 1 種の誤り（α）
本当は差がある	第 2 種の誤り（β）	正しい判断（$1-\beta$） [検定力]

2.3 サンプルサイズ（Sample Size）

検定は，手元にある，サンプル（標本，sample）のデータから，母集団（population）の平均値差を推定する推測統計の考え方を利用しており，サンプルサイズ（標本数）が大きくなればなるほど，統計的に有意である（$p < .05$）という結果になりやすいという大きな問題がある。このため，ある検定を行ったところ，20 人では有意ではなく，200 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。つまりサンプルサイズを大きくすれば検定力は高まるが，逆に，検定力が強すぎる場合には，実質的な差がなくても，有意な差を検出する可能性も高くなるのである。

このような観点から，村井（2006）は検定力分析（power analysis）を利用し，サンプルサイズを決定するいくつかの方法を提示している。検定力分析では，(a) 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（$1-\beta$）を確認する方法と，(b) 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（$1-\beta$）からサンプルサイズを決定する目的のものが多い（Field & Hole, 2003, p.154）。検定力分析はフリーソフトである G*Power 3（Faul,

Erdfelder, Lang, & Buchner, 2007）やデータ解析環境 R を使って実行することができる（豊田, 2009）。

2.4 効果量（Effect Size）

効果量とは,「効果の大きさ」のことを指し，実験的操作（experimental manipulation）の効果や変数間の関係の強さ（strength of association）を表す指標である（Field & Hole, 2003, p.152）。p 値は前述のように，サンプルサイズによって変わるものなので，実質的効果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が用いられることとなった。芝・南風原（1990）によると，効果量は,「測定単位にたよらない指標となっている。そのため，効果量を用いれば，単位の異なる変数を用いた研究の間でも，実験条件の効果の大きさを互いに比較することができる」(p.118) と定義されている。実験の条件によっては，有意差があっても（$p < .05$），実質的効果があまりない（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すべきである（American Psychological Association, 2001; Field, 2009; Kline, 2004 など）。

3. 本稿の目的

パラメトリック検定とノンパラメトリック検定は従来の検定で用いられてきたが，正確な p 値を得ることができる検定は現代のコンピュータ技術の進歩によって利用が増えてきている。そのため，本研究では，以下の 3 つを比較することを目的とした。

- ■ パラメトリック検定
- ■ ノンパラメトリック検定
- ■ 正確な p 値を得ることができる検定
  - (1) 平均値差の検定（並べ替え検定・確率化検定）
  - (2) クロス表の検定（正確確率検定）

4. 比較する検定の種類と概要

以下では，従来の検定で用いられるパラメトリック検定とノンパラメトリック検定，そして，近年注目されることが多い，正確な p 値を得ることができる，並べ替え検定（確率化検定）と正確確率検定の概要を説明する。

4.1 パラメトリック検定とノンパラメトリック検定

「パラメータ」とは，母集団の分布のことを指しているため，そのような母分布を想定している検定はパラメトリック検定と呼ばれている。例えば，*t* 検定や分散分析（ANOVA）の前提条件は以下のようなものである。

1. 標本の無作為抽出
2. 各グループの母集団の分布が正規分布（正規性）
3. 各グループの母集団の分散が等しいこと（等分散性）

ゆえに，パラメトリック検定は，母集団が正規分布し，母分散が等しいということが前提条件として挙げられる。サンプルサイズが小さい（例えば，*n* が 30 以下の）場合にも，「観測値の独立性，母集団分布の正規性，等分散性，といった検定導入のための統計的な前提が適切であれば、標本の大きさ（ただし *n* >= 2）に関係なく保証されるので、小標本で検定を行うのは自然」（前田忠彦，私信，2009 年 7 月 26 日）であると考えられている。さらに，パラメトリック検定はサンプルサイズが大きい場合には，正規性，等分散性の仮定に対しては，頑健性（robustness）を持つことが知られているため，検定の文脈においては最も使用頻度が高い。

それに対し，ノンパラメトリック検定では，母集団の分布に対して特定の仮定をせず，どのような母分布から無作為抽出されたサンプルであるかは問題としない（ただし，母集団からのランダムな標本抽出が前提)。用いる代表値は，順位・中央値などになる。ノンパラメトリック検定は，母集団の分布については仮定しないが，母代表値の推定を行っているため，推定という意味ではパラメトリック検定と同様に推測統計であり，サンプルから母集団への結果の一般化を目指している手法である。

ノンパラメトリック検定は，データがパラメトリック検定の前提を満たさない場合や，外れ値を含んでいてパラメトリック検定の結果に影響を与えていると考えられる場合に用いられる。2 群の場合はパラメトリック検定の *t* 検定（もしくは分散分析），3 群以上の場合は分散分析に相応する表 2 のようなノンパラメトリック検定がよく用いられる。

表 2　ノンパラメトリック検定の種類

群の数	データの対応	ノンパラメトリック検定
2 群	なし	マン・ホイットニーの *U* 検定
	あり	ウィルコクスンの符号付順位和検定
3 群以上	なし	クラスカル・ウォリスの順位和検定
	あり	フリードマン検定

Note. 前田（2004, p. 61）を基に作成

検定力は，パラメトリック検定とノンパラメトリック検定の両方で計算可能である。正規分布が仮定できるデータに対しては，パラメトリック検定の検定力が高く，外れ値を含む，歪んだ（正規分布ではない）データに対しては，ノンパラメトリック検定の検定力が高くなる。また，パラメトリック検定におけるデータの前提条件が満たされているデータに対して，ノンパラメトリック検定を行うと，検定力が低くなることが知られている（Field, 2009, p. 551）。具体的には，正規分布しているデータに対して，ノンパラメトリック検定を行うと，パラメトリック検定の検定力の 95.5%程度の検定力になると言われている[2]（Lehmann, 1975）。以上，パラメトリック検定とノンパラメトリック検定の比較をまとめたものが表 3 である。

表 3　パラメトリック検定とノンパラメトリック検定の比較

条件	パラメトリック検定	ノンパラメトリック検定
分布の仮定	正規分布	必要なし
等分散性	仮定	仮定
正規分布のとき	◎	○
外れ値が存在	×	○
$n < 6$	△	×

Note. 浜田（n.d.）を基に作成。$n < 6$ でノンパラメトリック検定が使えない理由は，有意水準 5%で有意になることがないためである（青木, 2009）。

4.2 「正確な *p* 値」を求める方法

本節では，「正確な *p* 値」を求める方法として，(a) 平均値差の検定として「並べ替え検定（確率化検定）」を，そして，(b) コーパス言語学研究で用いられるようなクロス表での頻度の検定として「フィッシャーの正確確率検定」の 2 つの説明を行う。

「正確な *p* 値」とは，ここで説明する 2 つの方法が，パラメトリック検定やノンパラメトリック検定のように，特定の確率分布を基に推定を行うわけではないので，*p* 値の計算において，母集団の未知のパラメータや，サンプリング誤差（sampling error）が入らないため，計算上も正しい *p* 値が得られる。そのため，「正確な（exact）*p* 値」と呼ばれる（Corcoran & Mehta, 2001, p. 4）。つまり，パラメトリック検定やノンパラメトリック検定で推定している *p* 値は，以下で説明する方法で得られる *p* 値の「近似（approximation）」を行うものであるため，正しい *p* 値は並べ替え検定（確率化検定）やフィッシャーの正確確率検定によって得られるものである。

これらの正しい *p* 値を求める方法は，比較的新しい方法として紹介されているが，実際

[2] *t* 検定とマン・ホイットニーの *U* 検定を比較した場合，*t* 検定の検定力×0.955 (=3/π [π は 3.1415…])で求められる。

は 1930 年代に提唱されており，考え方自体はかなり古いものである。しかし，莫大な計算を行うことができる高性能コンピュータが当時はなかった。その方法が近年，コンピュータの処理速度が飛躍的に速くなっているため利用が可能になってきている。

4.2.1 平均値差の検定（並べ替え検定，確率化検定）

平均値差の正しい p 値を求める方法は，並べ替え検定（permutation test）や確率化検定（randomization test）と呼ばれる[3]。本稿では「並べ替え検定」に名称を統一して以下，説明していく。並べ替え検定はノンパラメトリックな方法に分類される。大きな枠組みとしては，ブートストラップ法やジャックナイフ法とともに，手元の標本データを母集団と見なし，そこから無作為抽出を繰り返すことによって得られる情報を利用する，リサンプリング（resampling）の一種として紹介されることも多い。

ここでは具体的な例によって説明を加えていく。まず，図 1 中の左端表のような 2 群のデータが得られたとする。Group A と Group B はランダムに割りつけられており，平均値に 30.00 の差がある。並べ替え検定では，この 2 群はもともと同じ母集団に属していたサンプルで，母集団を代表している値であると考える（帰無仮説は「2 つの母集団は同じものである」）。よって，2 群の間に生じた 30.00 の差はグループへの割りつけによって偶然生じた差であると見なす。その場合に，この平均の差はどれぐらいの確率で起こるものなのかを計算する。そのために，図 1 のように，手元のデータ 12 個をプールし，そこから各群に 6 個ずつ再度割りつけるという作業を行う。

例：2 群のデータ(n = 12)

Group A	Group B
20	10
40	70
30	80
40	90
40	70
50	80
M = 36.67	M = 66.67
SD = 10.33	SD = 28.75

差 = 30.00

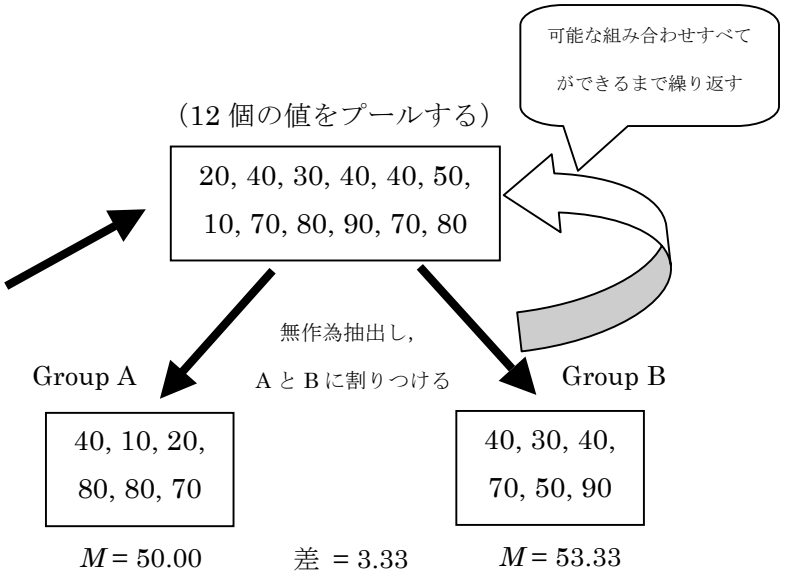

図 1　並べ替え検定のイメージ

[3] 実際は平均値だけではなく，どのような統計量でも対象とすることができる。

$n = 12$ で，各群に 6 個のデータが割りつけられる場合のすべての組み合わせは，12!/6!6! = 924 組ある（!は階乗を表す）[4]。この 924 組のうち，2 群の平均値の差が 30.00 かそれ以上（平均の差の絶対値で）になる組み合わせは 54 あり，54/924=0.0584 が p 値となる（両側検定）。この p 値が $p > .05$ であるため，「5%水準で有意差はない」という結果になる。

次に，同じデータを使ってパラメトリック検定とノンパラメトリック検定の比較を行った。t 検定をしてみると，$t(10) = -2.41$, $p = .037$，効果量 $d = 1.39$ となり，$p < .05$ で有意差があるという結果になる。この 2 群のデータの正規性や等分散性を確認してみると，Group B のデータは Shapiro-Wilk の正規性検定で正規性が確認できず，2 群の等分散性の仮定も満たされていない。そのため, Welch の検定を行ってみると $p = .051$ となる。また，ノンパラメトリック検定である，マン・ホイットニーの U 検定を行ってみると，$U = 6.00$, $z = -1.94$, $p = .052$, 効果量 $r = .62$ という結果になった。ちなみに，マン・ホイットニーの U 検定では，正確確率が計算できるため，確認してみると $p = .058$ となり，並べ替え検定と同じ p 値が算出されることが確認できる。これは並べ替え検定と同様に組み合わせの計算を行っているからである（橘，1997, p.55）。表 4 はこれらの検定結果の p 値比較のまとめである。

表 4　2 群のデータ例の検定結果 p 値比較（両側検定）

検定の種類	p 値
並べ替え検定	.058
t 検定（等分散を仮定）	.037
t 検定（Welch 検定）	.051
マン・ホイットニーの U 検定（近似）	.052
マン・ホイットニーの U 検定（直接確率）	.058

t 検定における検定力を G*Power3 で算出したところ, Power = .52 となり, Cohen (1988) の推奨する .80 にはかなり遠い値となっており，第 2 種の誤りを犯す確率が $\beta = .48$ であり高めになっている。非常に大きな効果量（$d = 1.39$）があるデータであるため，検定力を目標とする .80 を超えるようにするには，各群ともに 11 名（合計 22 名）にすれば良い（その場合の検定力 = .84）。このように，検定力分析によって，どれぐらいの検定力が得られたか計算したり，今後の実験に（もしくは実験前に）必要なサンプルサイズを見積もることもできるので，検定力分析は検定を行う場合には非常に有用である（詳細は，村井, 2006 や豊田, 2009 を参照）。

[4] Excel では FACT 関数を使い=FACT(12)/(FACT(6)*FACT(6))とセルに入力すれば計算できる。また，R でこの計算を行う場合は choose(12,6)で可能。各群のサンプルサイズが違う場合は，以下のようにして求めることができる。

```
library(combinat)     #パッケージの読み込み
nc<-nCm(9,5,4)        #9 に合計サンプル数，5，4 にそれぞれの群のサンプル数を入れる
nc                    #結果の出力
```

以上までで，さまざまな角度から 2 群のデータ例を分析してみたが，(a) 前提が満たされていないデータに対して t 検定を行うと結果が歪む，(b) 正しい p 値だけにこだわるのであれば，並べ替え検定が役に立つ，という 2 点のことがわかるだろう。

もう一度，並べ替え検定と従来のパラメトリック・ノンパラメトリック検定の違いを，t 検定を例に挙げて確認しておくと，t 検定によって得られる p 値は，並べ替え検定における p 値に近似させるために推定された値である。特にサンプルサイズが大きい場合，正規分布でなくても，中心極限定理（central limit theorem）のような特性によって，平均値の分布は正規分布に近づくため，t 検定によって並べ替え検定の近似が可能である。つまり，t 検定の結果と並べ替え検定の p 値が大きく違う場合は，常に並べ替え検定が正しく，そのデータは t 検定の前提を満たしていないということが主張できる（Hesterberg, Moore, Monaghan, Clipson, & Epstein, 2005, p. 51）。

並べ替え検定には 2 つの考え方がある。1 つは推測統計と同じく，母集団への一般化が可能であるという考え方である。この考え方の短所としては，並べ替え分布は手元のデータに依存するため，そのデータが（特にサンプルサイズが小さい場合には）本当に母集団を代表しているものなのかわからないということである。つまり，手元のデータが本当に母集団を代表しているものであると確信が持てるときのみ，そのような一般化が可能であると考えられる。

2 つ目の考え方は，「そのような母集団への一般化は考えていない」（橘, 1997）というものである。そもそも，従来の検定では（パラメトリック・ノンパラメトリック検定ともに）母集団から無作為抽出したサンプルを用いていないため，標本の無作為抽出という推測統計において，最も重要な前提が満たされていないため，その手法を分析に適用するのは厳密には誤用であるということを橘（1997）は強く指摘している。そして，推定の対象を母集団としている並べ替え検定（permutation test）と，結果は手元のサンプルにのみに適用することを目指している確率化検定（randomization test）とは，同じ計算をして，同じ p 値を得るが，目指しているところが違うとも述べている（p. 59）。ただし，思想的には根本的に違うものの，並べ替え検定であっても，確率化検定であっても，正確な p 値を得ることができるという事実は変わらない。ゆえに，並べ替え検定の利用者は，どこまで一般化するかを手元のデータの特性を考えて，分析をすればよいだろう。

このような一般化の問題に加えて，並べ替え検定では，サンプルサイズがかなり大きくなると組み合わせの数が膨大になり，計算量も多くなるという問題がある。しかし，並べ替え検定におけるリサンプリングは 999 回でよいという見解や（Hesterberg, Moore, Monaghan, Clipson, & Epstein, 2005, p. 54），実用的には「1 万回やれば充分」（粕谷, 1998, p. 170）という意見があり，最近のコンピュータではほぼストレスなく計算できるようになっているため，大きな問題ではないだろう。並べ替え検定は，R を使えば簡単に計算できる。

また，フリーで入手できる Excel のマクロを使う方法もある[5]。

4.2.2 クロス表の検定（フィッシャーの正確確率検定）

クロス表の検定として使用される，フィッシャーの正確確率検定（以降，正確確率検定）も，基本的な考え方は並べ替え検定に非常によく似ている（高倉, 2008）。具体的には，並べ替え検定の場合と同様に，あり得るすべての組み合わせを行い，サンプルに見られる統計量がどのぐらいの確率で起こるかを計算する。

クロス表の検定では，カイ 2 乗（χ^2 検定）がこれまでコーパス言語学の分野でも使用されることが多く，クロス表内の全期待度が 5 よりも小さい低頻度を扱う場合は，イェーツの連続性の修正（Yates' continuity correction）で χ^2 分布の近似度を上げるという方法がとられる（小林・田中, 2009）。その他にも，低頻度語がある場合には，対数尤度比検定/G 検定（log-likelihood ratio test/G-test）を用いることもあるが，これらの検定も，χ^2 分布による近似を行っているため，サンプルサイズが小さい場合には正確な p 値が得られない。

コーパス言語学において，正確確率検定を使う利点は 2 つ考えられる。まず 1 点目は，並べ替え検定の場合と同じく，従来の検定では算出することができなかった「正確な p 値」を正確確率検定によって得ることができるという点である。そして，2 点目は，コーパスにおけるサンプル（標本）と母集団の関係によるもので，例えば，夏目漱石の作品すべてをコーパスにした場合には，そこから得られる語の頻度は母集団の情報であり，夏目漱石の作品のサンプルとはならない（夏目漱石の他の作品はない）。さらに，その頻度情報を他の明治・大正時代の文豪作品への一般化を目指すためにサンプルとして使用するということもないだろう。つまり，コーパスは包括的なものであればあるほど，母集団とサンプルの区別がつかなくなる。そのような場合に，母集団・サンプルの区別なしに，正確な p 値を算出することができる正確確率検定は有効な方法であるといえる。

以下では具体的な例によって説明を加えていく。まず，表 5 のようなクロス表の頻度がイギリス英語とアメリカ英語のコーパスから，フレーズ A とフレーズ B について得られたとする。正確確率検定は，このクロス表においてデータが取り得るすべてのパターンを考え，表 6 のようにセル内の頻度を記号で表し，その確率を以下の式で計算する。

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

[5] 例えば（http://sci.kj.yamagata-u.ac.jp/~columbo/Stat/）など。このサイトにあるマクロは Excel2000 用であるが，Excel2003 や 2007 でも動く。並べ替え検定だけではなく，ほとんどのノンパラメトリック検定ができる，非常にすぐれたマクロである。

表 5　クロス表（2×2）の例

	イギリス英語	アメリカ英語	合計
フレーズ A	2	10	12
フレーズ B	8	5	13
合計	10	15	25

表 6　上記表 5 のセル内頻度を記号で示したもの

	イギリス英語	アメリカ英語	合計
フレーズ A	a	b	a+b
フレーズ B	c	d	c+d
合計	a+c	b+d	n

表 5 のクロス表の観測値の組み合わせに対して表 6 を当てはめて，この計算をすると確率（*p*）は .026 となる。すべてのパターン（3,268,760 通り）において，この値と同じもしくは小さな値が得られるパターンは 135,344 通りであるため，135,344/3,268,760 = 0.041 が正確確率検定によって得られる正しい *p* 値となる。

この例では，前述のように，期待値 5 以下のセルがあるので，カイ 2 乗検定を適用するのはふさわしくないが，比較のために表 7 のようにいくつかの検定を行った。その結果，カイ 2 乗検定や対数尤度比検定では，正確な *p* 値よりも低めの値が得られ，イェーツの連続性修正を行ったカイ 2 乗検定は，*p* = .060（有意差なし）という結果になった。「期待値が 5 以下のものがあればイェーツの連続性の修正を行う」という一般的に受け入れられている方法でも，カイ 2 乗検定と同じく「近似的な *p* 値」であるため，正確な *p* 値が問題になるときには正確確率検定のほうが良いということがわかる。

これらいくつかの検定の *p* 値の違いからも正確確率検定があることによって，サンプルサイズが小さい場合は特に，正確な *p* 値は何なのかを確認できるため実行する価値があるといえるだろう。

表 7　クロス表のデータ例の検定結果 *p* 値比較（両側検定）

検定の種類	*p* 値
フィッシャーの正確確率検定	.041
カイ 2 乗検定	.022
カイ 2 乗検定（イェーツの連続性修正）	.060
対数尤度比検定	.019

カイ 2 乗検定の検定力分析（α=.05）による検定力は .63，そして正確確率検定の場合は .59 であり，推奨されている .80 にはかなり遠い値となっていて，有意な差を検出する力が弱い検定であったことがわかる。また，効果量は Cramer の *V* = .46 と中程度であっ

た[6]。これら結果から得られた情報を利用して，G*Power 3 により，十分な検定力を得るにはどれぐらいのサンプルサイズが必要かを検討したところ，カイ 2 乗検定の検定力は図 2，正確確率検定の検定力は図 3 のようになった。これによって，カイ 2 乗検定，正確確率検定ともにサンプルサイズを 40 程度以上とすれば，目標とする検定力 .80 を超えることができるということがわかる。

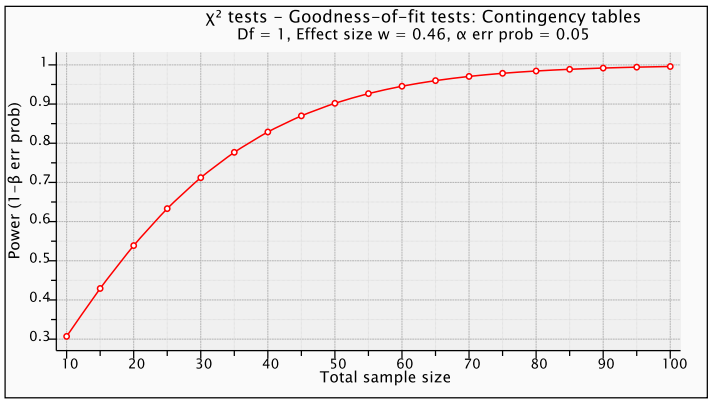

図 2　カイ 2 乗検定の検定力分析（必要サンプルサイズと検定力の関係）

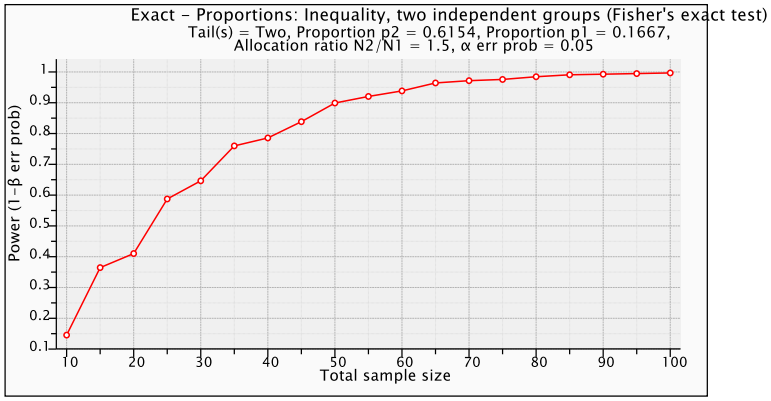

図 3　正確確率検定の検定力分析（必要サンプルサイズと検定力の関係）

[6] 2×2 のクロス表では，Cramer の V は φ 係数と呼ばれることもあり，ω とも数値の大きさが対応している（豊田, 2009, p. 101）。Field（2009, p. 699）はこれらの効果量指標の他に，2×2 のクロス表ではオッズ比（odds ratio）を提示すると解釈に役立つと主張している。

5. まとめ

本稿では，サンプルサイズが小さい場合の統計的検定について，従来のパラメトリック・ノンパラメトリック検定と，正確な p 値を得ることができる並べ替え検定・直接確率検定の比較を行った。その結果，正しい p 値が確認することができるという点で，並べ替え検定・直接確率検定による方法はまず初めに推奨されるべきであるということがわかった。

また，コンピュータの発達により，並べ替え検定や正確確率検定を行うのは無理ではなくなったので，わざわざ p 値の近似値を求める従来のパラメトリック検定やノンパラメトリック検定よりも，直観的かつ，わかりやすい結果が得られるといえるだろう。以下のコメントにあるように，今後，これらの手法はますます広がっていくと考えられる。

> “I believe that in a short time they will overtake what are now the more common nonparametric tests, and may eventually overtake the traditional parametric tests” (Howell, 2002, p. 692).

本稿では，検定方法の比較だけではなく，検定において重要な 4 つの要素の中から，問題なく研究者が考慮に入れているであろう，α とサンプルサイズの他に，コーパス言語学や外国語教育学の分野で，報告がまだ少ない効果量や検定力分析についても考察を加えた。p 値だけではなく，検定におけるこれらの重要な要素にも目を向けることによって，検定の精度を高めていき，より客観的な検定結果の判断と解釈が可能になると考えられる。

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 教授）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，小泉利恵 氏（常磐大学），印南 洋 氏（豊橋技術科学大学），および統計数理研究所共同研究プロジェクト「言語研究と統計」のメンバーの諸氏から貴重なアドバイスを頂いた。ここに記して感謝する。

文　献

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

青木繁伸 (2009).『フリードマン検定とウィルコクソン符号付き順位検定について』Retrieved from http://aoki2.si.gunma-u.ac.jp/taygeta/statistics.cgi?mode=res&no=11168

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.

Corcoran, C. D., & Mehta, C. R. (2001). Exact level and power of permutation, bootstrap and asymptotic tests of trend. Retrieved from http://www.cytel.com/Papers/monteboot.pdf

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/aap/projects/gpower/

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

Field, A., & Hole, G. (2003). *How to design and report experiments*. London: SAGE.

浜田知久馬 (n.d.).『数理統計学(第十回)ノンパラ検定とは？1』 Retrieved from www.rs.kagu.tus.ac.jp/hamada/file/Class/stat11.ppt

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., & Epstein, R. (2005). Bootstrap methods and permutation tests. In D. S. Moore & G. P. McCabe (Eds.), *Introduction to the practice of statistics* (5th ed., pp. 11–70). New York: W. H. Freeman.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/Thomson Learning.

粕谷英一 (1998).『生物学を学ぶ人のための統計のはなし』東京:文一総合出版.

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

小林雄一郎・田中省作 (2009).「χ 二乗検定再入門」『英語コーパス学会第 33 回大会ワークショップ資料』

Lehmann, E. L. (1975). *Nonparametrics. Statistical methods based on ranks*. San Francisco, CA: Holden-Day.

前田啓朗 (2004). 「少人数学級での差の検討―ノンパラメトリック検定―」. 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. 53–62) 東京 : 大修館書店.

村井潤一郎 (2006). 「サンプルサイズに関する一考察」 吉田寿夫 (編) 『心理学研究法の新しいかたち』(pp. 114–141). 東京 : 大修館書店.

芝 祐順・南風原朝和 (1990).『行動科学における統計解析法』東京:東京大学出版.

橘 敏明 (1997).『確率化テストの方法―誤用しない統計的検定―』 東京:日本文化科学社.

高倉耕一 (2008).「統計解析におけるコンピュータの新しい利用法:確率化テスト, モンテカルロ法」『生活衛生』*52*, 221–228.

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.