特別寄稿

順天堂大学医療看護学部　医療看護研究23
P.1－8(2019)

サンプルサイズ諸論

Discussions on Sample Size

戸ヶ里　泰　典[1)]
TOGARI Taisuke

Ⅰ．はじめに

CONSORT声明やSTROBE声明に代表されるように、この20年の間に様々な臨床研究や疫学研究の質保証のためのガイドラインが提示されてきている。サンプルサイズの決定方法についてはこうしたガイドラインにおいて共通して必ず挙げられている項目の一つである。また、研究倫理指針の中にもサンプルサイズの設計が挙げられており、大学院生をはじめとする研究初心者は研究計画の立案段階でまず直面する問題になっている。

サンプルサイズの設計は、量的研究の科学的水準の担保に繋がる要素の一つであり、研究実施においてきわめて重要である。それもあって研究計画や研究方法に関するクリティークではよく指摘される部分でもある。しかし、「サンプルサイズはどのように計算したか」と一言指摘することは容易いが、実際のところ研究デザインや研究目的とも絡んでくるために、一意では決めることはできず、いくつかの要素を複合して決定していくことになる。逆に「サンプルサイズ」というキーワードを通じて推測統計手法や研究の方法論を眺めるといろいろ見えてくることがある。具体的なサンプルサイズの決め方については別の専門書に譲るとして、本稿ではサンプルサイズが推測統計手法や調査研究方法論にどのようにかかわっているのかについて様々な観点から整理をし、サンプルサイズについて考えると同時に統計的仮説検定を中心に統計解析の手法を再度見直す手がかりを提示していきたい。

Ⅱ．統計的仮説検定とその問題点

１．統計的仮説検定の歴史

統計的仮説検定は、1925年にロナルド・フィッシャー（Sir Ronald A. Fisher 1890-1962）の書“Statistical Methods for Research Workers（研究者のための統計的方法）”において示された統計手法である（Salsburg et al., 2001/2006）。フィッシャーは統計的手法の普及に尽力し、いわゆる有意水準P=0.05の提唱者とも言われている。つまり、この本の中で、“The value for which P = .05, or 1 in 20, is 1.96 or nearly 2 ; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not.”と述べており、これがその後統計手法を用いる各領域の研究者の間に広まっていったとされる（Bangdiwala, 2016）。必ずしもフィッシャーはカットオフ値として提唱をしたのではないが、フィッシャーによる提唱ののちに、イエジー・ネイマン（Jerzy Neyman 1894－1981）とエゴン・ピアソン[i]（Egon Pearson 1895－1980）によって推測統計学（抽出した部分集団から母集団における特徴や性質を推測する統計学）の方法として確立したものとされた。

帰無仮説や対立仮説という名称はネイマンとピアソンによって命名され、フィッシャーの検定を有意性検定として区別されてもいる（土居, 2010）。例えば、(1)ネイマン＝ピアソンは検定仮説と対立仮説を設定するが、フィッシャーは検定仮説のみを設定する。(2)ネイマン＝ピアソンではサンプルの大きさは事前に定められるが、フィッシャーにとってそれは重要でない。(3)ネイマン＝ピアソンは仮説の採択・棄却の判定を機械的に反復せよとしているが、フィッシャーでは一回限りの判定でよいとされる、など、相違がみられており、フィッシャーの死まで両者間では論争が絶えなかった（フィッシャーより一方的にネイマンとピアソ

1) 放送大学
The Open University of Japan

i カイ２乗検定や積率相関係数の提唱者であるカール・ピアソン（Karl Pearson）の息子である。

ンに批判がなされていた）ようである（Salsburg et al., 2001/2006）。現代の統計的仮説検定は、ネイマンとピアソンにより定式化され、ネイマン=ピアソンの理論、あるいはネイマン=ピアソンの仮説検定論、と呼ばれ、初等数学のテキストにおいても登場し広く浸透している。本稿では、このネイマン=ピアソンの統計的仮説検定を扱っていく。

２．統計的仮説検定の手続き

統計的仮説検定においては次の順序で作業が行われる。(1) 帰無仮説と対立仮説を立てる、(2) 検定統計量と分布を決める、(3) 有意水準と棄却域を決める、(4) データを取得し検定統計量を算出する、(5) 仮説の棄却と採択を行う、のそれぞれである。

帰無仮説とは、検定において棄却する、つまり否定されるべき仮説のことを指す。主に「値に差がない」とか「値が０である」という仮説が立てられる。対立仮説とは研究者が正しいとしたい仮説のことで、「値に差がある」とか、「値が０でない」という仮説が立てられる。

次に、検定統計量とは、帰無仮説が正しいとするための確率Pを計算するための統計量で、一般にサンプルのデータから計算される。なお、帰無仮説の設定によって、検定統計量が従う分布が異なる。例えば、ｔ検定を行う場合はＴ分布、カイ二乗検定を行う場合は、χ^2分布、分散分析を行う場合はＦ分布を用いる。

有意水準とは、帰無仮説を棄却する範囲を決める水準のことで、「α」とも記される。この水準より低い確率であれば、帰無仮説は棄却できるとする。通常は５％に設定するが、先に述べたように、必ずしも根拠が明確にある数値ではない。また、５％というカットオフ値による二分法的な考え方が、この統計的仮説検定の本質ではあるが、後で述べるように昨今ではこうした考え方に疑義が呈されていることも念頭に置く必要があろう。

以上の準備のもとで、サンプル（分析対象者）からデータを取得し、検定統計量を算出し、帰無仮説が従う分布に基づいて生起確率を割り出し、棄却・採択を判断することになる。

３．統計的仮説検定に対する様々な批判

統計的仮説検定に対しては、様々な批判が挙げられてきた。大久保と岡田（2016）は問題点を整理している。ここでは大きく２つの問題点を紹介する。

一つ目は論理上の問題点で、後件肯定の誤謬という論理学的問題を抱えているとされている点である。つまり、有意差があるならばP<.05であるという前提があって、次に、P<.05であることがわかった時、それが即ち有意差がある、とは必ずしも言えないということである。例えば、「キツネザルであるならば霊長類である」という前提があった時に、考古学者が見つけ出したある生物の化石が「霊長類である」ことがわかった時、それが即ちキツネザルといえるかどうかはわからない。なお、後件否定は真であることがわかっている。例えばその化石が「霊長類でない」ことがわかった時に、それはすなわちキツネザルでないということができる。（ただしこの論理については、確率論には適用できないとも言われている。）

なお、統計的仮説検定はこの誤謬を表面的に解消するために対立仮説を設け、その支持においては背理法を用いている。もし有意差がないとすると、P<.05つまり帰無仮説の棄却という結果は説明できない。したがって、有意差はある、という論理になる。このように遠まわり論理での検証をしており、本来研究者がみたい差があるという仮説が正しいという確率そのものを見ているということではない。この点について各所から疑問が呈されてきている。

二つ目の解釈における問題点とは、「有意差」の解釈に対する問題で、実質的に重要な差であるかどうか、という点である。つまり、統計的仮説検定における検定結果は、有意水準を設定し、その水準の範囲を有意とし、範囲外では有意とはならないとする二分法による考え方を有している。このことは、P=.049とP=.051とでは決定的な相違があることを意味するが、それが実質的に意味のある相違であるといえるのかという点である。また、有意水準ではなく有意確率そのものに目を向けたとしても、有意確率は研究者が関心を持つ実質的な差や関連性の大きさなどの効果そのものではなく、あくまでも検定統計量に基づく確率を見ているに過ぎない。

４．統計的仮説検定におけるサンプルサイズの問題

統計的仮説検定では、検定統計量の算出における手続き上、サンプルサイズが大きくなるとP値[ii]は小さくなる。例えば、相関係数の検定（ｔ検定）を見ていくと、r_{xy}がｘとｙの相関係数、Ｎをサンプルサイズと

ii 学術雑誌によって大文字P、大文字イタリックP、小文字ｐなどと表現が異なるが本稿では大文字イタリックPと表現する。

表１　サンプルサイズと*P*値の関係：相関係数=0.2の検定（t 検定）の場合

N	t 値（t_{xy}）	*P*値
50	1.414214	0.163753
100	2.020726	0.046036
200	2.872281	0.004519
300	3.523729	0.000492
400	4.072264	0.000056
500	4.555217	0.000007
1000	6.448514	0.000000

表２　サンプルサイズと*P*値の関係：差の平均＝5.0、不偏分散＝100.0の対応のあるt検定の場合

N	t 値（t_d）	*P*値
50	1.414214	0.163753
100	2.000000	0.048268
200	2.828427	0.005159
300	3.464102	0.000610
400	4.000000	0.000076
500	4.472136	0.000010
1000	6.324555	0.000000

すると、

$$t_{xy} = \frac{|r_{xy}|\sqrt{N-2}}{\sqrt{1-r_{xy}^2}}$$

である。たとえば相関係数が0.2の時、**表１**のように*P*値は変化する。

また、対応のあるｔ検定の場合、ケースごとの差の平均M_d、差の不偏分散σ^2（σ：シグマ）、サンプルサイズをNとすると、

$$t_d = \frac{M_d}{\sqrt{\frac{\sigma^2}{N}}}$$

となる。例えば差の平均=5.0、不偏分散=100.0とすると、**表２**のように*P*値は変化する。

このように同じ相関係数や平均の差であってもサンプルサイズが大きくなるほど有意確率は小さくなっていく性質を持っている。つまり有意確率だけで結果を判断することはいささか危ういことがこのことからも伺われる。

５．統計的有意性と*P*値に関するASA声明

2016年に*P*値の適正な使用と解釈の基礎にある原則を明確にする目的で、アメリカ統計協会（American Statistical Association：ASA）から声明“The ASA's Statement on p-Values：Context, Process, and Purpose”が出された（Wasserstein et al., 2016）。この中では**表３**に示す骨子が提示されている。

表３　統計的有意性と*P*値に関するASA声明の骨子

- *P*値はデータと特定の統計モデルが矛盾する程度を示す指標の一つである
- *P*値は調べている仮説が正しい確率や、データが偶然のみで得られた確率を測るものではない
- 科学的な結論や、ビジネス、政策における決定は*P*値がある値を超えたかかどうかのみに基づくべきではない…統計的有意性は科学的結論を主張するための保証として広く用いられているが科学のプロセスを著しく損ねている
- 適正な推測のためには、すべてを報告する透明性が必要である…データ収集の際のすべての決定（サンプルサイズ含む）、すべての統計解析、すべての*P*値
- *P*値や統計的有意性は、効果の大きさや結果の重要性を意味しない。
- *P*値はそれだけでは統計モデルや仮説に関するエビデンスのよい指標にはならない…*P*値以外のアプローチが適切かつ実施可能な場合は*P*値を計算しただけでデータ解析を終えるべきではない

日本計量生物学会から日本語訳（http://biometrics.gr.jp/news/all/ASA.pdf）が提示されており、本稿はそこから引用した。

このように、*P*値はあくまでもいくつかある推測統計指標の一つとして位置づくのであって、唯一絶対的な根拠とするものではない。ASA声明では、*P*値を補う､あるいは別のアプローチの採用を推奨している。特に、推定（信頼区間、予測区間、信用区間）中心の報告、ベイジアンアプローチ、False Discovery Rateなどが例として挙げられている。

ただし、このASA声明は統計学的仮説検定を完全否定するということではない。問題点はあるものの統計学的仮説検定が一世紀近く支持され使用され続けてきている積極的な意義もある[iii]。そうした意義を生かしつつ*P*値が持つ問題点をある程度補う方法として検定力分析が提唱されている。

Ⅲ．検定力分析とサンプルサイズ

１．効果量とは

効果量とは、群間での平均値の差の程度、変数間の関連の強さなど、研究関心の程度を表す値を、データの単位に左右されないよう標準化したものを指す。統

iii 例えば確率という共通する基準で解析結果を評価することが可能であること、統計学的有意であるというある意味クリアな解析結果を踏まえて考察をすることが可能であること、などが挙げられるだろう。

計的仮説検定における帰無仮説が正しい時は効果量がゼロとなる。帰無仮説が正しくない場合は、帰無仮説からの乖離の程度に応じて大きくなる値である。昨今では、分析結果には効果量を報告することが必要とされている（Chavalarias et al., 2016）。

効果量と検定統計量とは大きく関係がある。一般に検定統計量は、N（サンプルサイズ）の関数とes（効果量）の関数の積で表される（大久保ら，2016）。

$$検定統計量 = f(N) \times g(es)$$

つまり検定統計量のうち、サンプルサイズによらない部分が効果量であるといえる。したがって、同じ効果量であってもサンプルサイズが大きくなればなるほど統計量は大きくなり、有意確率が小さくなる。なお効果量は、大きくd族（d family）とr族（r family）に分けられている。

d族効果量は、主に平均値の差の効果量を表し、(母)平均の差÷(母)標準偏差で計算される。母平均は標本平均を使用できるが、母分散（母標準偏差）は標本分散を調整する必要がある。最もよく知られており、使用されている指標は、独立した２群の差の効果量であるCohenのd（Cohen's d）である。ここで、グループ１と２のサンプルサイズをn_1, n_2、グループ１と２の平均値M_1, M_2、グループ１と２の標本分散をS_1^2, S_2^2、とすると

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$$

$$Cohenのd = \frac{M_1 - M_2}{S}$$

となる。このほかにも、不偏分散を使用したHedgesのgや、gをさらに補正したGlassのΔ（デルタ）、対応のあるデータの場合はdやΔのほか、分母の標準偏差を両者の共分散で補正したd_Dなどが知られている。

r族効果量の例は、ピアソンの相関係数rが最もよく知られている効果量である。また、回帰分析のときは決定係数（R^2）をもとに算出するf^2という値を用いる場合もある。

$$f^2 = \frac{R^2}{1 - R^2}$$

分散分析のときはη（イータ）２乗を用いる。ここで、偏差平方和とは、(値－平均)^{2}を足したものであり、級間偏差平方和＋級内偏差平方和＝全体の偏差平方和となる（詳細は統計学のテキストなどを参照されたい）。これらを用いて以下の式で計算される。

$$\eta^2 = \frac{級間偏差平方和}{全体の偏差平方和}$$

多要因分散分析の際には偏η^2（η_p^2）を用いる。この際、関心のある要因、関心のない要因、誤差の３つの変動に分かれるため、全体の偏差平方和＝関心要因の偏差平方和＋それ以外の要因による偏差平方和＋誤差平方和となり、以下のように計算できる。

$$\eta_p^2 = \frac{関心要因の偏差平方和}{関心要因の偏差平方和 + 誤差平方和}$$

なお、母分散の推定にこだわった形でη^2を補正したω^2, ω_p^2（ω：オメガ）も使用されることがある。

ノンパラメトリック検定における効果量は、２×２表の時はϕ（ファイ）係数、m×n表のときはクラメールの連関係数Vが相当する。

２．効果量の解釈

効果量は相対的な数値であり、明確な基準はないが、Cohenは経験的な解釈として、次のように示している（Cohen, 1969）。小さな効果量とはd＝0.2、中程度の効果量とはd＝0.5、大きな効果量とはd＝0.8である。また、他の効果量と併せて**表４**に示す目安を示している（Cohen, 1988）。しかし、小さな効果量でも重要な意味を取りうる場合もあるため、先行研究や関連する研究を踏まえて、最終的には研究者が判断する必要がある。

３．２種の誤りと検定力

統計的仮説検定において、第１種の過誤とは、帰無仮説が正しいのに帰無仮説を棄却してしまう過誤を指す。つまり、「差がない」のに「差がある」としてしまう過誤で、この過誤の確率が有意確率となる。第１種の過誤の確率はαエラーとも呼ばれることに起因してギリシャ文字でαと書かれる。統計的仮説検定では通常α＝0.05（５％）に設定される。

他方、第２種の過誤とは、帰無仮説が正しくないのに帰無仮説を採択してしまう過誤を指す。これは「差

表４　Cohen（1988）による効果量の解釈の目安

	効果量	効果小	効果中	効果大
平均の差のt検定	d	0.20	0.50	0.80
相関	r	0.10	0.30	0.50
カイ二乗検定	W	0.10	0.30	0.50
分散分析	η^2	0.01	0.06	0.14
回帰分析	R^2	0.02	0.13	0.26
	f^2	0.02	0.15	0.35

がある」のに「差がない」とする過誤であり、βエラーとも呼ばれるため、この過誤の確率はギリシャ文字でβと書かれる。なお、統計的仮説検定ではあくまでもαについて扱っており、第２種の過誤βについては関心がもたれていない

このβの逆である、帰無仮説が誤っているときに帰無仮説を棄却できる確率１－βは検定力（検出力）と呼ばれている。一般に0.8が設定されることが多く、α＝0.05と併せて５－80ルールとも呼ばれている(Cohen, 1988)。αとβはトレードオフ関係にあることから、効果量や有意水準が一定とすると、検定力が高い場合サンプルサイズは大きくなる、という関係がある。これらの中で事前に研究者が能動的に操作できるものがサンプルサイズである。こうした検定力・効果量・有意水準・サンプルサイズは互いに影響しあうという関係を用いて、事前にサンプルサイズを決定する際に検定力分析(power analysis)が行われる。なお、この分析について、統計ソフトRのpwrパッケージにおいて実施できる。また、専用無料ソフトも出ており[iv]、定評があるG*Power（http://www.gpower.hhu.de/）が良く使われている[v]。

４．２つの検定力分析

データ収集前に、サンプルサイズを決める目的で、推測される効果量、有意水準、検定力からサンプルサイズを算出する分析が行われる。その一方で、事後の検定力分析と呼ばれる分析もある。データ収集、分析後に検定力を確認する目的で、サンプルサイズ・算出された効果量、有意水準から検定力を算出することが行われる。主に帰無仮説が棄却されなかったときに実施される。

G*Power分析例として、今回は独立したサンプルのｔ検定を実施する際の例を示す（**図１**）。

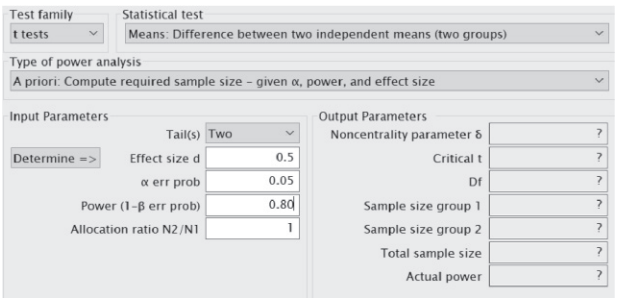

図１　G*Powerの入力画面

iv医学系に特化したソフトPS: Power and Sample Size Calculationも知られている。http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize#Windows

v G*powerなどの検定力分析ソフトの使い方については、国内外のウェブサイトにとてもわかりやすい資料が大量に落ちているので参考にするとよい。

はじめに、Test familyおよびStatistical testの部分で検定の種類を選ぶ。今回は２つの独立したサンプルのｔ検定を選択した。Type of power analysisでは、今回は事前分析なのでA priori（事前分析）を選択し、Tails（両側か片側か）では、Two tails（両側）を、効果量（effect size）dの大きさとしては0.5（Cohenの基準では効果量が大の値）を設定し、有意水準（α err prob）は0.05、検定力（power 1－β err prob）は0.80、サンプルサイズの比（Allocation ratio N2/N1）は１：１として入力を行った。計算（calculate）させると次の**図２**の画面になる。

右下の部分に計算結果が表示され、この場合各グループ64名の計128名（Total sample size）と表示がなされた。

Ⅳ．推定とサンプルサイズ

１．検定と推定

検定とはこれまでに見てきたように、抽出したサンプル（標本）から母集団（population）における母数（パラメータ）に関する仮説を検証する方法であった。こ

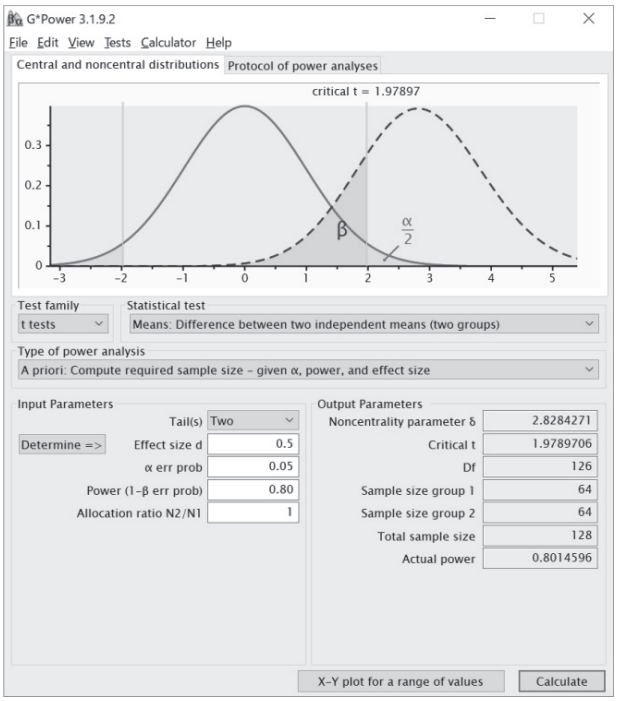

図２　G*Powerの計算結果画面

こで母数とは、母集団の特徴を指す。主に平均値、分散、相関係数、回帰係数、などが相当する。その一方で同様に母集団における特徴について推測する推測統計の手法の一つに推定がある。推定とは抽出したサンプルから母集団における母数を推定する方法のことで、大きく点推定と区間推定の２種類がある。

点推定とは、サンプルのデータから計算された１点の値を母集団の点推定値とするもので、母数の種類によって異なる。例えば、平均値はそのまま点推定値となるが、分散は不偏分散が点推定値となる。相関係数や回帰係数はそのまま点推定値[vi]となり、効果量や比率はそのまま点推定値となる。

２．信頼区間

ただし、点推定だけで真の母数を表現するには限定的で、区間推定を用いることで表現をする方がより真の母数の表現につながる。この区間推定の考え方は1934年にイエジー・ネイマン（Neyman J）によって提案された。ネイマンは「信頼区間（confidence interval：CI）」と呼び、信頼区間の両端を「信頼限界」と呼んだ（Salsburg et al. 2001/2006）。慣例で信頼水準を95％とする場合がほとんどである。95％が指す意味については、次のように説明できる。例えば小学６年生の平均身長の点推定値をｍとし標本抽出を繰り返すとし、100回抽出した時（100個サンプルができたとき）に100個信頼区間ができる。この100個の信頼区間のうち95個の信頼区間にｍが含まれると期待して良い、という意味である。

信頼区間の求め方は母数の種類によって異なる。平均値の場合サンプルサイズが小さい場合は必ずｔ分布を用いる。しかし、大きい場合（25以上とも30以上とも）標準正規分布を用いても良いとされる。標準正規分布で求めるとき、ある平均値の点推定値ｍの場合、SE（standard error）：標準誤差として、

$$m - 1.96 \times SE < m < m + 1.96 \times SE$$

となる。

なお、標準誤差とは標本平均が母平均に対してどのくらいばらついているのかの程度を表す。

３．信頼区間とサンプルサイズ

平均値の標準誤差SEの値を算出する式には、分母にルートN（サンプルサイズ）があった。これはNが大きくなればなるほど標準誤差は小さくなることを意味している。例えば平均値50、標準偏差10の時の95％信頼区間について示す（**表５**）。なお、N=50以上は標準正規分布を使用している。

表５　サンプルサイズ（N）と信頼区間との関係

N	95%CI下限	95%CI上限	区間の幅
5	45.53	54.47	8.9443
10	46.84	53.16	6.3246
50	48.59	51.41	2.8284
100	49.00	51.00	2.0000
500	49.55	50.45	0.8944
1000	49.68	50.32	0.6325
5000	49.86	50.14	0.2828
10000	49.90	50.10	0.2000
50000	49.96	50.04	0.0894
100000	49.97	50.03	0.0632

ここでサンプルサイズが100倍になれば区間の幅は10分の１になることがわかる。信頼区間の幅は広いほど精度が悪く利用価値が少ないと評価される。幅が狭いほど精度が高く利用価値が高い。しかし、幅が狭くなりすぎると信頼度が下がり区間内に数字が収まる確率は下がる。したがって、信頼区間の幅をどの程度に収めたいか、研究者は検討が必要で、逆にこの幅を定めることによって、サンプルサイズを求めることが可能になる。

平均値の95％信頼区間は

$$m - 1.96\sqrt{\frac{s^2}{N}} \leq m \leq m + 1.96\sqrt{\frac{s^2}{N}}$$

信頼区間の幅をδ（デルタ）以下とすると

$$\delta \geq m + 1.96\sqrt{\frac{s^2}{N}} - m + 1.96\sqrt{\frac{s^2}{N}} = 2 \times 1.96\sqrt{\frac{s^2}{N}}$$

$$N \geq \frac{2^2 \times 1.96^2 \times s^2}{\delta^2}$$

となり、信頼区間の幅をどの程度にしたいのか（δ）を定めることを通じてサンプルサイズを探索していくことができる。信頼区間の求め方については、母数の種類によって様々である。他については統計学のテキストを参照されたい。

４．統計的仮説検定と信頼区間の関係

帰無仮説が推定値＝０、対立仮説：推定値は０でない、という１標本ｔ検定の例を考える。この時、信頼区間に帰無仮説の値（＝０）が含まれていないと、

viただし不偏分散を用いて計算する。

統計的仮説検定の結果帰無仮説は棄却されることになる。これは検定の方法と信頼区間の式とが関係があるためである。

例えば、推定値mで標準偏差s、N（= 6）のとき信頼区間は

$$m - 2.57\frac{s}{\sqrt{N}} \leq m \leq m + 2.57\frac{s}{\sqrt{N}} \leftrightarrow$$

$$m - 2.57SE \leq m \leq m + 2.57SE$$

となり、１標本$T = \frac{m}{\frac{s}{\sqrt{N}}} = \frac{m}{SE} \sim t_5$である。

このとき、５％水準で自由度５のｔ分布の両側$\alpha / 2$のT= ±2.57となり、m≥|2.57SE|の時に５％水準で帰無仮説は棄却される。m≥|2.57SE|の時は信頼区間には０は含まれないことからわかる。ロジスティック回帰分析結果のオッズ比の信頼区間に１を含まないと統計学的有意である、という事実も、対数を用いている関係で１になっているだけで、同様の理由による。

Ⅴ．多変量解析とサンプルサイズ

多変量解析では、扱う変数が多くなり、それに応じて推定する母数の数が増える。少なくとも推定する母数の数よりも多いサンプルサイズでないと計算ができないという問題がある。さらに、解析結果の安定性を鑑みて、50＋８×説明変数数、あるいは、104＋説明変数数という式（Green, 1991; Tabachnick et al., 2007）も提唱されている[vii]。

また、構造方程式モデリングの場合100または200以上（Boomsma, 1985）、推定パラメータあたり５または10ケース（Bollen, 1989）、など経験的ではあるが、様々提案されている。また、“A-priori Sample Size Calculator for Structural Equation Models”（http://www.danielsoper.com/statcalc/calculator.aspx?id=89）という計算サイトもあり、Westland（2010）による論文の基準に基づいてサンプルサイズ計算をすることができる。

二項ロジスティック回帰分析のサンプルサイズについては、シミュレーションの結果、従属変数のいずれか少ないカテゴリのサイズが説明変数×10以下であると、結果のバイアス、精度、モデルフィット等問題が生じていることが示された。しかし、少ないカテゴリのサイズが説明変数×10以上では問題がなかったことが示された（Peduzzi et al., 1996）。この論文のインパクトは大きく、昨今でもロジスティック回帰分析のサンプルサイズは説明変数×10以上が必要とされている場合が多い。

なお、看護・保健系の研究で多く見られるモデル探索的な量的研究の場合、検証すべき仮説が明確にはない場合もある。この際には明確なサンプルサイズの基準はなく、サンプルサイズは大きいほど良いとも言われている（高木ら, 2006）。したがって、事前に欲しい効果サイズ、あるいは信頼区間の幅が決まっていないこうした探索的研究では事前分析はできない。仮説を明確に打ち出している研究のタイプでのみ、サンプルサイズの計算ができる。ただし、明確な仮説がない研究であっても多変量解析を実施する研究ではサンプルサイズの設計が必要である。自身の研究のタイプ・特性を慎重に踏まえて検討をしていく必要がある。

Ⅵ．まとめ

本稿では統計的有意検定の問題点について整理し、それと併せて検討が必要な推測統計の方法について概観しつつ、サンプルサイズについて様々な角度から整理をしてきた。サンプルサイズは、その決定には大きく３つの方向性があった。ひとつは検定力分析によって算出する方向性であり、帰無仮説検定の結果は、有意水準、検定力、サンプルサイズ、効果量の４つの要素の関連性から生じることに基づいた算出方法であった。この方法については多くの無料ソフトが出ており、本稿ではG*Powerというソフトの使用方法について説明を行った。二つ目は、区間推定に基づく方向性であった。推定したい母数の95%信頼区間の幅の事前設定により、必要なサンプルサイズを定めることができるという方向性である。この２つの方向性は、解析結果に関する仮説が明確に定められている研究であることが前提となっていた。三つ目は、多変量解析の実施に伴う方向性である。これは解析結果の安定性に関するシミュレーションの結果から、様々な基準が提案されていた。

サンプルサイズの設計に際しては必ずしもこれら３点のすべてを踏まえる必要はなく、またひとつも踏まえることができない研究もありうる。研究の目的やデザインに即した形で柔軟に設計をしていくことが必要である。

[vii]これらはエフェクトサイズ$\beta \geq .20$, 有意水準 $\alpha \leq .05$, 検定力80%がベースとなっている

文献

Bangdiwala SI. (2016). Understanding significance and p-values. *Nepal J Epidemiol*. 522-524. doi：10.3126/nje.v6i1.14732.

Bollen K. (1989). Structural Equations with Latent Variables. John Wiley. New York.

Boomsma A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, 50, 229-242.

Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. (2016). Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA - J Am Med Assoc*. 315(11), 1141-1148. doi：10.1001/jama.2016.1952.

Cohen J. (1969). *Statistical Power Analysis for the Behavioral Science*. Academic Press. New York.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Science (2nd Eds.)*. Lawrence Erlbaum Associated. Hillsdale.

土居淳子(2010). 帰納的推論ツールとしての統計的仮説検定：有意性検定論争と統計改革. 年報人間関係学, 13, 15-36.

Green S. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behav Res*, 26, 499-510.

大久保街亜, 岡田謙介 (2012). 伝えるための心理統計. 勁草書房.

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstem AR. (1996). A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*, 49(12), 1373-1379. doi：10.1016/S0895-4356(96)00236-3.

Salsburg D. (2001/2006). 竹内惠行, 熊谷悦生訳, 統計学を拓いた異才たち：経験則から科学へ進展した一世紀. 日本経済新聞出版社.

Tabachnick BG, Fidell LS. (2007). Using Multivariate Statistics. Pearson/Allyn & Bacon. Boston.

高木廣文, 林邦彦 (2006). エビデンスのための看護研究の読み方・進め方. 中山書店.

Wasserstein RL, Lazar NA. (2016). The ASA's Statement on p-Values：Context, Process, and Purpose. *Am Stat*, 70(2), 129-133. doi：10.1080/00031305.2016.1154108.

Westland JC. (2010). Lower bounds on sample size in structural equation modeling. *Electron Commer Res Appl*, 9, 476-487.