シリーズ臨床心理学研究と統計学

4. サンプルサイズと検定力分析

井　上　俊　哉
Shunya INOUE

以下では，２つの独立な平均値差のｔ検定（両側検定）を例に話を進めるが，ここでの議論は，ほかの検定の場合にも当てはまる。

1. サンプルサイズが小さいと何が問題なのか?

10人の被験者を各群５人ずつに分けてｔ検定をしたと聞いたら，「サンプルサイズが小さい（=被験者数が少ない）な」という印象を持つかもしれない。では，サンプルサイズが小さいと何が問題なのだろうか?

1.1 正規分布からの逸脱に対する頑健性

サンプルサイズが小さいことの不都合の１つとして，検定の頑健性に依拠しにくくなることが挙げられる。頑健性(robustness)については，このシリーズの第２回(井上, 2005[1])で触れたが，「想定されたモデルが必ずしも正しくない場合に，そのモデルを仮定して適用された統計的手法のモデルの逸脱に対する鈍感さを表す概念」である（芝ほか，1984[2]）。Glass and Hopkins(1996)[3]は，正規性，等分散，独立というｔ検定の重要な３つの仮定について，仮定からの逸脱がｔ検定に与える影響を調べた研究結果を概観し，たとえば正規性について，「両側検定の場合にはほとんど気にしなくてよく，片側検定の場合でも小さい方の群で20以上のサンプルサイズがあれば頑健である」とまとめている。もちろん，母集団分布が確かに正規分布であれば，サンプルサイズが小さくてもｔ検定を用いることに原理的な問題はないのだが，現実の研究では母集団分布が正規分布であることを確信できない場合が多く，頑健性の議論はありがたい。（等分散，独立からの逸脱に対する頑健性については，井上[1]，Glass & Hopkins[3]などを参照のこと）

サンプルサイズが小さいことの，さらに大きな問題は，検定における第２種の誤りの確率が大きくなることである。

1.2 検定における２種類の誤り

仮説検定を勉強したことのある人ならば，検定では「第１種の誤り（Type I error)」と「第２種の誤り（Type II error)」という２通りの誤りを犯しうることを，学んでいると思う（表1)。

第１種の誤りの確率，すなわち，H_0が真であるときに誤ってH_0を棄却する（「２群の母集団平均が等しいのに，平均値差が有意であると結論する」）確率は，検定を用いる人がかならず設定する有意水準αにほかならない。したがって，この誤りの確率は，サンプルサイズとは無関係に，5%

表1 検定における2種類の誤り

	H_0が真	H_0が偽
H_0を採択	正しい決定	第2種の誤り
H_0を棄却	第1種の誤り	正しい決定

教養部情報処理研究室

や1%といった小さな確率に保たれる。

第2種の誤りの確率，すなわち，H_0が偽であるときに誤ってH_0を採択する（「2群の母集団平均に差があるのに，平均値差は有意でないと結論する」）確率は，αに対してβで表される。研究では，母集団平均に差があると信じ，検定によってそれを実証しようと望むことが多いから，βを小さくすることは，研究者にとって重大な関心事でなければならない。ところが実際には，検定に際して有意水準αを報告することは常識なのに，βはあまり（まったく?）意識されない。その最大の理由は，「αを決めなければ検定ができないのに対して，βを知らなくても検定できてしまう」ことにあると思われる。しかも，αは研究者が自ら決めればよいが，βを求めるには面倒な計算が必要である。第2種の誤りの確率βは，サンプルサイズや設定された有意水準によっても変化するので，βを求めるには，「μ_1とμ_2の差はどの程度なのか」「サンプルサイズはいくつなのか」「有意水準は何%なのか」を考慮しなければならないのである（表2）。

1.3 サンプルサイズとβ

ここで，シミュレーションによって，βの大きさを概算してみよう。現実場面では母集団の平均を知ることはできない（だからこそ検定によって結論を導こうとする）が，シミュレーションでは母集団平均などの条件を既知のものとして設定し，その条件下で多数回の実験を繰り返し，t統計量の分布の様子などを確かめることができる。ここで考えるシミュレーションは以下のようなものである（[シミュレーション1]）。

> 2つの学習条件における母集団の平均をそれぞれ40点と42点，母集団の標準偏差をともに8点に設定し，それぞれの母集団から5名ずつのサンプルを抽出してtを計算することを50,000回繰り返す。そして，有意水準5%のt検定（両側）を行うとき，第2種の誤りを犯す割合を求める。

図1中のヒストグラムは，シミュレートされた50,000個のtの分布，曲線は自由度8(=5+5-2)のt分布である。仮説検定では，H_0が真である（母集団平均が等しい）ことを前提として，自由度n_1+n_2-2のt分布上で採択域と棄却域を決める。今の例では，自由度8のt分布をもとにして，-2.31から2.31の範囲が採択域となる。ヒストグラムのうち，この採択域に含まれる部分の割合がβをシミュレートした値と考えられるが，図1においてその割合は0.93378にもなる。つまり，[シミュレーション1]の状況下では，本当は母集団平均に差があるのに，90%以上の確率で，検定の結果は有意にならない。

今の例では2群各5人とサンプルサイズが小さかったが，サンプルサイズを大きくすれば，βの値を小さくすることができる。ほかの条件は変えずに各群のサンプルサイズを50人ずつに増やした[シミュレーション2]の結果は，図2

表2 αとβの比較

α	β
第1種の誤りの確率	第2種の誤りの確率
仮説を棄却(「平均値に有意差あり」と結論)したときに，犯しているかもしれない誤り	仮説を採択(「平均値に有意差なし」と結論)したときに，犯しているかもしれない誤り
H_0のもとでのtの分布(t分布)上で求められる	H_1のもとでのtの分布(非心t分布)上で求められる
検定をする人が決める(計算は不要)	効果量，サンプルサイズ，αの大きさによって変化(計算が必要)
5%や1%といった小さな確率に保たれる	場合によっては，とても大きな確率になる

の通りで，この場合のβは，0.76526 にまで小さくなっている(それでも，まだかなり大きいが)。

以上から，つぎのことがわかる。

・サンプルサイズが十分でないとき，第 2 種の誤りを犯す確率は非常に大きいかもしれない。

・サンプルサイズを大きくすれば，第 2 種の誤りを犯す確率を下げることができる。

2. 有意水準と効果量の影響

サンプルサイズだけではなく，「有意水準」および「μ_1 とμ_2の差」も，第 2 種の誤りの確率の大きさに影響する。

2.1 有意水準とβ

まず，有意水準αと第 2 種の誤りの確率βの関係を見てみよう。有意水準を変えるということは，棄却域・採択域の大きさを変えるということである。有意水準を 5%から 1%へと小さくすると棄却域が狭まり，第 1 種の誤りを犯す確率は小さくなるが，採択域は広がって，第 2 種の誤りを犯す確率βは大きくなってしまう（図 1 または図 2 を見ながらイメージしてもらいたい)。βを小さくするには有意水準を大きくすればよいのだが，代償として第 1 種の誤りの確率αは大きくなってしまう。

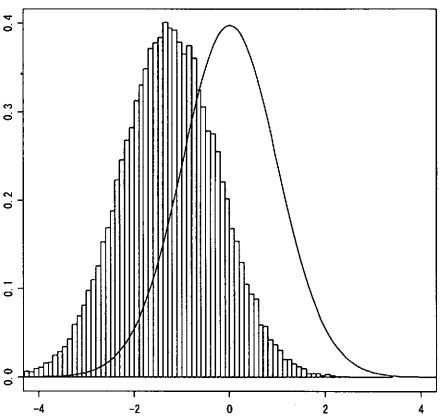

図2 μ_1=40, μ_2=42, σ=8, n_1=n_2=50

・第 1 種の誤りの確率αと第 2 種の誤りの確率βの間には，一方を小さくすると他方が大きくなるという関係がある。

2.2 効果量とβ

つぎに，「μ_1 とμ_2の差」とβの関係について，シミュレーションで確かめてみよう。2 群の母集団平均を，40 と 42 のかわりに 40 と 48 に変えて，サンプルサイズは各群とも 50 名で［シミュレーション 3］を行った結果が図 3 である。有意水準 5%のときのβは 0.00178 で，「μ_1 とμ_2の差」のほかは同じ条件だった［シミュレーション 2］と比べて格段に小さくなっている。

2 つのシミュレーションの比較から，μ_1 とμ_2の差，すなわち$\mu_1-\mu_2$ の絶対値が大きいほど，βが小さくなることが示されるが，$\mu_1-\mu_2$の大きさには，測定単位に依存するという問題がある。たと

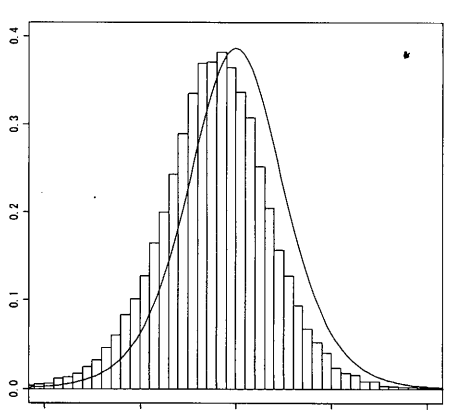

図1 μ_1=40, μ_2=42, σ=8, n_1=n_2=5

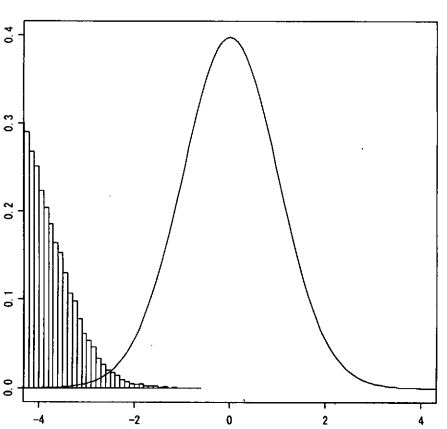

図3 μ_1=40, μ_2=48, σ=8, n_1=n_2=50

えば，被験者を2つの学習条件に無作為に割り当てたのち同じ課題に取り組んでもらい，学習に要した時間を測定するという実験を考えよう。2群の所要時間の母集団平均の差が30秒であるとする。つまり，測定単位を「秒」にとるかぎりは，$\mu_1-\mu_2=30$ である。ところが，30秒は0.5分と表すこともでき，測定単位を「分」にすると，$\mu_1-\mu_2=0.5$ になってしまう。このように，実質は全く同じ差であるのに，測定単位次第で，$\mu_1-\mu_2$ の大きさは変わる。この例のように測定対象が時間であれば，測定単位をすべて「秒」に換算することで単位を共通化することも可能だが，多くの心理学的尺度には，「分」や「秒」あるいは「cm」「kg」などの絶対的な単位が存在しないため，揃える単位を決めることは事実上不可能である。

この問題は，$\mu_1-\mu_2$ のかわりに，$\mu_1-\mu_2$ を2群に共通の母集団標準偏差σで割った $d=\frac{\mu_1-\mu_2}{\sigma}$ を用いることで解決できる。d は「効果量」「標準化された平均値差」あるいは「Cohen の d」とよばれ，測定単位に依存しない。時間の例でいえば，同じデータを「秒」で表した場合と「分」で表した場合とでは，標準偏差の大きさが60:1になるため，30秒も0.5分も効果量に換算すると同じ値になる。われわれのシミュレーションに関しては，最初の2つでは，$\mu_1=40$，$\mu_2=42$，$\sigma=8$ であったから効果量は（(40-42)/8=）-0.25，最後の例では $\mu_1=40$，$\mu_2=48$，$\sigma=8$ であったから効果量は（(40-48)/8=）-1.00になる。βの大きさを効果量の関数として定式化すれば，測定の標準偏差を気にすることなく，以下のように一般化することができる。

・効果量の絶対値が大きいほど，第2種の誤りを犯す確率は低くなる。

効果量はそれ自体とても興味深い指標であり，このシリーズの第1回でも簡単に触れた（井上，2004[4]）が，さらに次回で詳しく取り上げる予定である。

3 検定力分析

3.1 検定力

ここまで，サンプルサイズ，有意水準，効果量と第2種の誤りの確率との関係を概観してきたが，統計学の分野では，1-βを検定力（power）とよび（検出力と訳されることもある），βよりも1-βとサンプルサイズ，有意水準，効果量の関係を議論することが多い。

検定力，サンプルサイズ，有意水準，効果量の4つの値の間には，どれか3つが決まると残りの1つが決まるという関係があり，この関係を分析することは検定力分析とよばれる。検定力分析は，①有意水準，効果量，サンプルサイズの関数としての検定力；②効果量，有意水準，検定力の関数としてのサンプルサイズ；③有意水準，サンプルサイズ，検定力の関数としての効果量；④サンプルサイズ，検定力，効果量の関数としての有意水準の4種類を考えることができる（Cohen，1988[5]）（表3）が，役立つことが多いのは①あるいは②の検定力分析である。①の検定力分析を行えば，すでに行われた検定（あるいはこれから

表3 4種類の検定力分析

x ⇒ y	G*Power 3における名称
有意水準，効果量，サンプルサイズ ⇒ 検定力	Post hoc
効果量，有意水準，検定力 ⇒ サンプルサイズ	A Priori
有意水準，サンプルサイズ，検定力 ⇒ 効果量	Sensitivity
サンプルサイズ，検定力，効果量 ⇒ 有意水準	Criterion

行おうとする検定）について，有意水準，サンプルサイズ，研究で期待されている効果量の見積りから，検定力を算出できる。

研究の計画段階では，効果量を見積り，有意水準を設定し，確保したい検定力を決めた状態で，必要なサンプルサイズを決めるという②の検定力分析がとくに重要である。他の条件が同じならばサンプルサイズが大きいほど検定力は高くなるが，サンプルサイズは多ければ多いほどよいとは，一概に言えない。むやみに多くの被験者を集めることは，労力，時間，必要経費のいずれの観点からもコストが高くつく。また，無理な収集はデータに偏りをもたらすおそれがあるし，実質的にほとんど意味のない微少な効果量であっても高い確率で有意になることは，必ずしも望ましいことではない。したがって，研究を始める前に，適切なサンプルサイズを知ることの価値は高い。

3.2 効果量の見積り

②の検定力分析における最初の難関は，効果量の見積りである。効果量概念に慣れないと，いくつになるのか，まるで見当がつかないだろう。効果量を見積もる上で知っておいて損のないポイントを3つ挙げておく。

- 母集団の平均値差（$\mu_1-\mu_2$）および2群に共通と仮定される母集団の標準偏差σについて見当をつけられるならば，式$\frac{\mu_1-\mu_2}{\sigma}$によって，大雑把な効果量を求めることができる。
- 近年，英文誌では効果量の推定値を報告する論文が増えており，複数の研究から得られた標本効果量を1つの値に統合するメタ分析も普及している。関心のある変数を含む先行研究が存在するならば，効果量を見積もる上で役に立つ。
- Cohen(1988)は，効果量の大きさについて，小さい効果量＝0.2；中程度の効果量=0.5；大きな効果量＝0.8という目安を与えている。多くの心理学研究における効果量の中央値は約0.5になるという報告もある（Lipsey & Wilson, 2001[6]）。

3.3 有意水準の決定

すでに触れたように，αとβの間には相反する関係がある。研究では，目的に応じてαの値が先に決められる。慣習上，αは10%，5%，1%のいずれかに設定されることが多いが，この中では10%にとると検定力はもっとも高くなり，1%にとると検定力はもっとも低くなる。同じ検定力を確保するには，α=1%のときに大きなサンプルサイズが必要になる。

3.4 検定力の決定

有意差が得られたときに「母集団平均に差があることが確認された」として議論を進めるために，仮説検定においては，第1種の誤り（母集団平均に差がないのに有意差があると結論する誤り）の確率を小さくすることが非常に大事になる。第2種の誤りも，「誤り」であるからには確率を小さくすべきであるが，αと比べるとやや大きめに設定されることが多い。よく用いられるのは，検定力80%（第2種の誤りの確率20%）という数字である。有意差を検出する確率を高めたい場合には，検定力を90%，95%などにすればよいが，その分大きなサンプルサイズが要求される。

3.5 適切なサンプルサイズの算出

以上が決まれば，いよいよサンプルサイズの計算である。手順について，芝・南風原(1990)[7]の10章や永田(2003)[8]などが参考になるものの，適用するのはなかなか難しい。Cohen(1988)[5]に検定力分析の結果が表として載せられているので，これを用いる手もあるが，一番のおすすめはフリーソフトG*Powerの利用である（2007年2月現在における最新バージョンはG*Power3。

http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/)。

G*Power は英語版だがとても使いやすく，先に挙げた 4 種類の検定力分析のすべてを簡単に実行できる（表 3 参照)。今回のシミュレーション 2 の状況（α=0.05，d=−0.25，$n_1=n_2$=50）を例にとり，Post hoc 分析で計算してみると，検定力は 25%足らず（0.235780）であり，この計画では有意差を得る確率は十分でないことがわかる。では，検定力を 80%まで上げるにはサンプルサイズをいくつにすればよいか? この問いに答えるために，A Priori 分析を行うと，2 群のそれぞれに 253 名という，かなり大きなサンプルサイズが必要であることが，たちどころにわかる。

引用・参考文献

[1] 井上俊哉 2005 ティ検定の頑健性 東京家政大学附属臨床相談センター紀要 第5集，pp. 91−97.

[2] 芝祐順・渡部洋・石塚智一 1984 統計用語辞典 新曜社.

[3] Glass,G.V., & Hopkins,K.D. 1996 *Statisitical Methods in education and psychology 3rd ed.* Allyn & Bacon.

[4] 井上俊哉 2004 平均値差をめぐって 東京家政大学附属臨床相談センター紀要 第 4 集，pp. 69−74.

[5] Cohen,J. 1988 *Statistical power analysis for the behavioral sciences. 2nd ed.* Lawrence Erlbaum Associates.

[6] Lipsey,M.W., & Wilson,D.B. 2001 *Practical meta-analysis.* Sage.

[7] 芝祐順・南風原朝和 1990 行動科学における統計解析法 東京大学出版会.

[8] 永田靖 2003 サンプルサイズの決め方 朝倉書店.