〔東京家政大学　臨床相談センター紀要　第5集　p. 91～97, 2005〕

シリーズ臨床心理学研究と統計学

2. t検定の頑健性
── t検定を使える条件 ──

井 上 俊 哉
Shunya INOUE

1. t検定

1.1 tの標本分布

2群の平均値差に関するt検定は、もっともよく使われる統計手法の一つである。t検定では、2群のサンプルサイズ、平均、分散から、t統計量を計算する。

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (1)$$

そして、計算されたtの値を自由度n_1+n_2-2のt分布に照らして、平均値差の有意性を判断する。tの計算式や自由度は、もちろん適当（でたらめ）に決められたものではなく、（一定の前提条件の下で）tの標本分布が自由度n_1+n_2-2のt分布に従うことが理論的に明らかにされている[1)]。

現実に標本を多数回繰り返し抽出して、理論を経験的に確かめるのは困難（事実上、不可能）であるが、コンピュータを使えば（コンピュータの性能と計算時間の制約の範囲内で）、特定の母集団からの標本を、数万回でも数十万回でも抽出することができる。そして、抽出した各標本についてtを計算し、その分布をグラフで表せば、H_0のもとでのtの標本分布を経験的に示すことができる（このように、現実世界における実験が困難、ないし不可能な現象について、できるだけ現実を擬した実験をおこなうことをシミュレーションと呼ぶ）。図1は、平均と分散が等しい2つの正規母集団のそれぞれからサンプルサイズ10の標本を50,000回繰り返し抽出して、50,000個のtを計算し、その分布をヒストグラムで示したものである。図1には、自由度18（=10+10-2）のt分布も重ねて描いてあるが、tのヒストグラムが自由度18のt分布に非常によくフィットしていることがわかる。（本稿におけるシミュレーションおよび作図は、近年注目されつつあるRというフリーの統計解析パッケージを用いている[1]）。

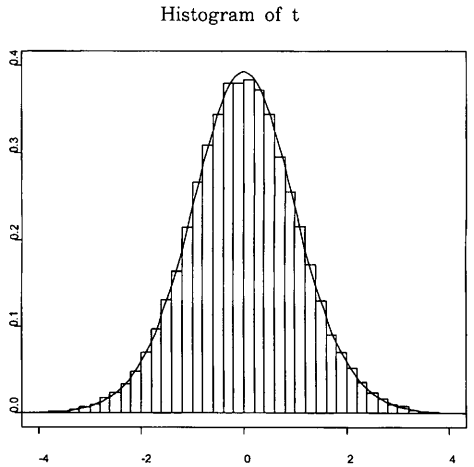

図1　50,000回計算されたtのヒストグラムと自由度18のt分布

1.2 t検定の前提条件

ところで、t分布の発明者ゴセットは、t統計量とその自由度を導出する際に、①2つの母集団

教養部 情報処理研究室

分布が正規分布である（正規性）、②2つの母集団分布の分散が等しい（等分散）、③2つの標本が2つの母集団から独立に抽出されている（独立）、という3つの条件を前提とした（図1のシミュレーションは、3条件をすべて満たしている）。

では、これら3つの条件のいずれかが満たされないとき、どのようなことが起こるのだろうか。ここでは、等分散の仮定が満たされない場合のシミュレーションを示そう。2つの正規母集団の分散を1：4に設定して、各母集団からサンプルサイズ16と4の標本を独立に抽出することを50,000回繰り返すシミュレーションを行い、得られたtのヒストグラムを描くと図2のようになる。この場合のヒストグラムは、図2に描かれている自由度18（=16+4-2）のt分布曲線と大きく離れており、自由度18のt分布にもとづいて有意水準5％の限界値を設定すると、tが限界値である±2.10を超える確率が、5％よりずっと大きくなることを読み取ることができるであろう。t検定に際してα=0.05に決めるということは、「平均値差がないときに有意差ありと結論する誤り（第一種の誤り）について、その確率を5％に設定します」と宣言していることを意味する。5％と宣言しておきながら、誤りの確率が実際にはこれよりもずっと大きくなっているとしたら、意図せずとも、虚偽の研究報告をしていることになる。

ゴセットの前提した3条件が満たされない場合に生じうる問題の一つは、このように、名目のα（設定した有意水準）と実際のα（設けられた棄却域に、計算されるtが落ちる確率）が食い違う可能性のあることである。よく知られているように、t検定には、(1)式で示される「等分散を仮定した場合のt検定」のほかに、「等分散を仮定できない場合に用いるt検定（ウェルチの検定）」が用意されている。ウェルチの検定は、名目のαと実際

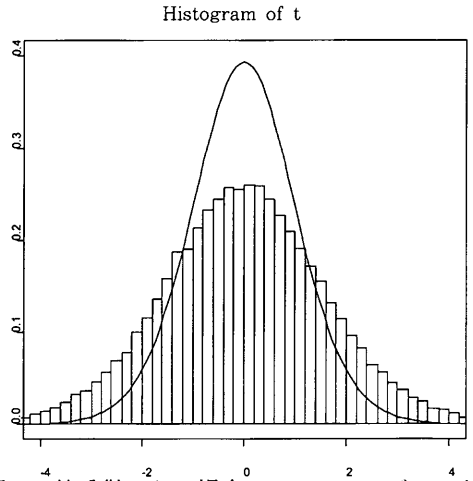

図2　等分散でない場合のtのヒストグラムと自由度18のt分布

のαの食い違いを小さくするための数学的工夫の成果である。等分散の仮定からの逸脱の場合だけでなく、独立や正規性の仮定が満たされない場合にも、代替の方法が存在する。独立の仮定が満たされない場合には、対応のある2群のためのt検定を用いることが常套手段であり、正規性が満たされない場合には、正規分布に近づける変数変換後のt検定の適用、あるいはノンパラメトリック検定（マン・ホイットニー検定）の採用が選択肢となる。

2. 頑健性

2.1 頑健性とは何か

上に見たように、条件が満たされない場合について、代替の方法も用意されてはいるのだが、一方で、統計学者は、t検定の頑健性についての研究も進めてきた。『統計用語辞典[2]』によれば、頑健性（robustness）とは、「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。ここでの文脈に即していうと、t検定は上記3つの仮定のもとで導

出されたものではあるが、たとえば、母集団分布が正規分布でない場合にも、t検定を用いた結論の正しさが損なわれないならば、t検定は正規性からの逸脱に対して頑健であるということになる。

t検定の頑健性については、これまでにもシミュレーション研究がおこなわれており（たとえば、Boneau, 1960[3]; Sawilowsky & Blair, 1992[4]）、一定の結論が得られている。日本の統計の教科書にはあまり紹介されていないが、米国の代表的な教科書（Glass & Hopkins[5], pp.290-296）には、以下のようなことが記されている。

(1) 正規性については、サンプルサイズがある程度以上大きければ頑健（とくに両側検定の場合）。
(2) 等分散については、両群のサンプルサイズが等しい場合に頑健。
(3) 独立については、区別することが重要。独立でない場合には、対応のある2群のためのt検定を用いるべきである。

本稿では、3つの仮定を逸脱するケースを何通りかシミュレートし、有意水準を両側5％に設定したとき、実際のαが名目のα(5%)とどの程度食い違うのかを確認していく。頑健性についてしっかり議論するには、第一種の誤りの確率αだけでなく、第二種の誤りの確率βも考慮しなければならないが、ここでは、第一種の誤りに関する頑健性にかぎって検討をする。

2.2 シミュレーション

正規性

正規性を逸脱した母集団分布としては、左に歪んだ分布、右に歪んだ分布、2峰以上の分布、一様分布など、さまざまな形状を想定できる上に、2群の母集団分布の形状が異なるケースも考えると、正規性からの逸脱には無数のバリエーションがある[2)]。ここでは、2群の母集団分布が同一の指数分布$f(x)=2e^{-2x}$（図3）である場合を母集団

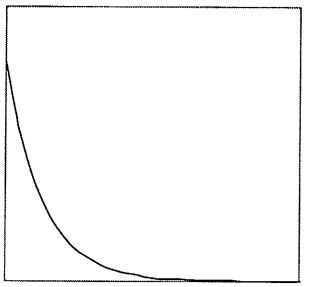

図3　指数分布

分布として設定した。2つの指数母集団から独立に各サンプルサイズ10の標本を抽出して50,000回のtを求めた結果のヒストグラムが図4である。重ねて描かれた自由度18のt分布への当てはまりもよく、50,000個のtのうち、5％の棄却域に落ちる割合は0.04372となった（このシミュレーションを何度も繰り返すと、割合は変動するが、0.04328、0.04378、…のように、名目上の値との差は1％未満の範囲であった）。サンプルサイズが小さいときには、名目のαと実際のαのズレは大きくなり、サンプルサイズを大きくすると実際のαは5％に近づいていく（各群5名で

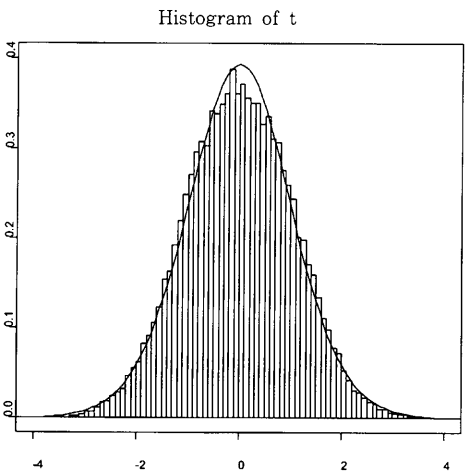

図4　正規性が満たされない場合のtのヒストグラム（$n_1=10$, $n_2=10$）

シミュレーションを繰り返すと、経験的αは0.03804、0.03866、0.03886、…；各群40名でシミュレーションを繰り返したときの経験的αは0.04822、0.04808、0.04828、…となる)。

2群のサンプルサイズが等しいとき、tの分布は左右対称であった(図4)が、2群のサンプルサイズが異なる場合(ここでは$n_1=16$、$n_2=4$)、計算されたtの分布は歪んでくる(図5)。しかし、両側5%の棄却域に落ちるtの割合を計算すると、0.04514、0.0448、0.0457、…となり、実際のαと名目のαとの違いは、それほど大きくない。

正規性からの外れ方は多くの場合分けが必要であり、これだけで一般的な結論を導くことは到底できないが、過去のシミュレーション研究の結果を総合すると、(1) 2群のサンプルサイズがほぼ等しく、(2) 各群25～30名以上の被験者があり(歪みが極端でない場合には各15名程度で十分)、(3) 両側検定の場合には、正規性からの逸脱は第一種の誤りに大きな影響を与えない。ただし、母集団分布がひどく極端に歪んでいて(とくに2集団の歪みの方向が逆の場合)、2つのサンプルサイズが偏っているときには、頑健とはいえないことも示されている(図5からも推察されるように、片側検定では実際のαが名目のαと大きく異なるおそれが強くなる)。そのような場合には、ノンパラメトリック検定(マン・ホイットニー検定)の使用を考慮すべきである。

等分散

図2では、2つの母集団分散が等しくない($\sigma_1^2:\sigma_2^2=1:4$)ときに、各母集団から大きさ16と4のサンプルを抽出すると、tのヒストグラムが自由度18のt分布よりも大きく広がっていることを見た(図6は図2を再び示したものである)。このケースについて、50,000回計算されたtのうち5%の棄却域(両側)に落ちるものの割合

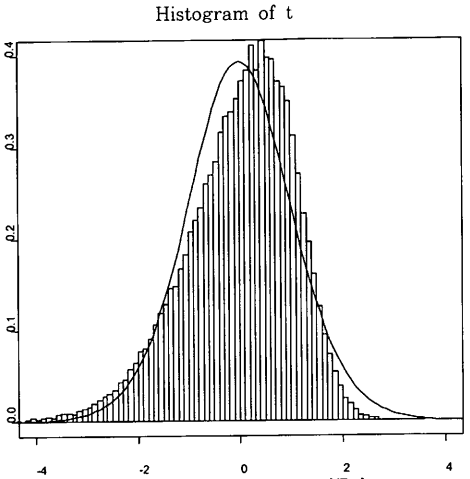

図5　正規性が満たされない場合のtのヒストグラム($n_1=16$, $n_2=4$)

を何度か求めると、0.18664、0.1891、0.18712、…のようになり、実質のαが名目のαを大きく上回っていることが確認される。

等分散の仮定からの逸脱の影響の仕方は、2群のサンプルサイズの比によって大きく変わる。同じく母集団の分散比が1：4でも、サンプルサイズが4と16のときには、図2の場合とは逆に棄却域に落ちるtの割合は、名目の値よりもずっと小さくなる(図7。この場合の実質のαは0.00612、0.00636、0.00682、…となる)。また、2つのサンプルサイズが等しい場合($n_1=n_2=10$)、等分散の仮定を大きく逸脱しているにもかかわらず、実際のαは名目のαと大きくは異ならない(図8。経験的αを求めると0.0548、0.05388、0.05468、…となる)。

図6から図8に示した3つのシミュレーションの結果から、サンプルサイズの比が母分散の比と逆方向のときには実際のαが膨らむこと、サンプルサイズの比と母分散の比が同じ方向のときには実際のαが小さくなること、サンプルサイズの比が1に近づくと2種類のαの食い違いが小さくなることを読み取ることができよう。2群の母分散

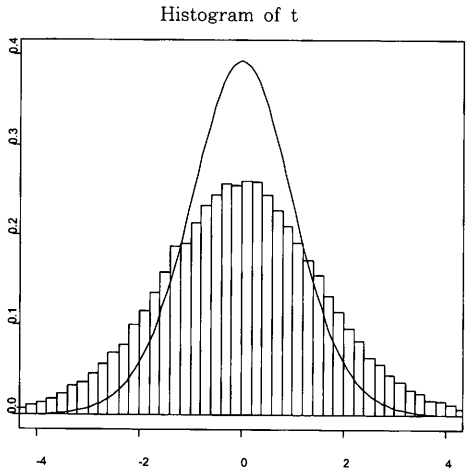

図6　母集団分散比が1：4場合のtのヒストグラム（$n_1=16$, $n_2=4$）

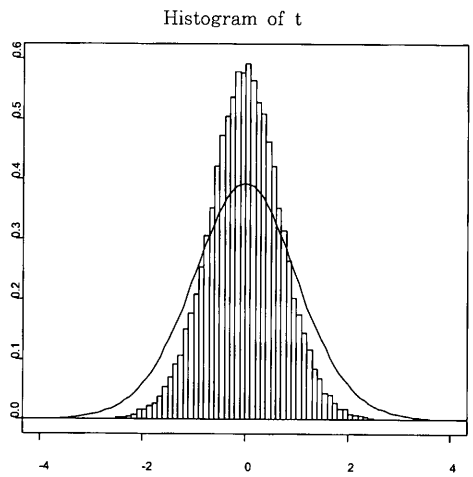

図7　母集団分散比が1：4場合のtのヒストグラム（$n_1=4$, $n_2=16$）

比とサンプルサイズ比がともに1から離れている場合には、第一種の誤りの確率が宣言した有意水準どおりにならないことに注意する必要がある。

一方、等分散を仮定しないウェルチの検定を用い、母集団分差比が1：4のケースで、いろいろなサンプルサイズについて、実際のαをシミュレーションしてみると、$n_1=16, n_2=4$のときには、0.06054、0.06108、0.06048、…、$n_1=4$, $n_2=16$のときには、0.0501、0.05056、0.0517、…、$n_1=10$, $n_2=10$のときには、0.05084、0.05028、0.0504、…となり、サンプルサイズが偏っていてもその影響が小さく、実質上のαと名目のαが近くなることがわかる。

これらの結果をまとめると、

(1) 2群の母分散が異なっていても、2群のサンプルサイズが等しければ、通常のt検定を使っても、実際のαは名目のαとほとんど変わらない。

(2) 2群の母分散が異なっていて、しかも2群のサンプルサイズが等しくないときに、通常のt検定を用いると、実際のαが名目のαと離れた値になる。

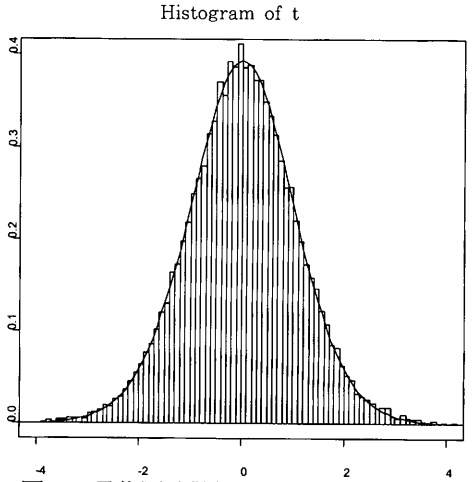

図8　母集団分散比が1：4場合のtのヒストグラム（$n_1=10$, $n_2=10$）

(3) ウェルチの検定を用いれば、2群の母分散比、サンプルサイズ比にかかわらず、実質のαが名目のαと比較的近くなる。

これだけを見ると、サンプルサイズによらず名目のαと実質のαが近くなるウェルチの検定の方が望ましく思われる。にもかかわらず、統計学の教科書で、(1)式のt検定が詳しく説明されるのはなぜだろうか。永田[6] (pp.187-188) は、以下のようなウェルチの検定の欠点を挙げ、平均値差

の検定を行う際には、できるだけサンプルサイズを揃えた上で、t検定を用いることが望ましいと解説している。

(1) 近似的な検定である。

(2) 自由度の計算が面倒である。

(3) 母分散が等しいとき、通常のt検定に比べると検定力が低下する。

正規性にせよ等分散にせよ、母集団に関する仮定であり、その仮定が満たされているかどうかに関して断定的に結論することはできない。仮定が満たされているのにノンパラメトリック検定やウェルチの検定を適用すると検定力が低下する。したがって、制約が少ないという理由だけで、それらの検定法を安易に採用することは慎まなければならない。2群のサンプルサイズを揃えさえすれば、仮定を気にせず通常のt検定を用いてよい(頑健である)というのは、ありがたい性質というべきであろう。

一方、サンプルサイズが大きく異なる2群のデータをもとにt検定を行わざるをえない場合には、正規性や等分散の仮説について慎重に吟味する必要があり、逸脱の程度が極端だと思われるならば、代替の検定の採用を検討すべきである(等分散の仮定について、永田[6] (p.194) は、サンプルサイズの比が2倍以上異なり、2つの標本分散が2倍以上異なる場合には、ウェルチの検定を用いるべきである、という目安を示している)。

独　立

対応のあるデータで各群を変数と見なすと、2変数間には正の相関が生じると考えられるが、そのようなケースでは、(1)式で求められるt統計量の分布は0付近に集中してくる。母集団相関を0.3に設定して求めた50,000個のtのヒストグラムが図9である(正規性、等分散は満たされているとし、サンプルサイズは各群とも10)。ヒストグラムは、自由度18のt分布よりも中心方向に集まっており、αの実際値が名目値よりも小さくなることは明らかであろう。対応のあるデータをとるとき、2変数間でこれより大きな相関が生じることも珍しくなく、その際には、実際のαと名目のαのズレは、もっと甚だしくなる。したがって、2群の間に対応がある場合には、(1)式にもとづくt検定を用いずに、対応のある2群のt検定を用いなければならない。

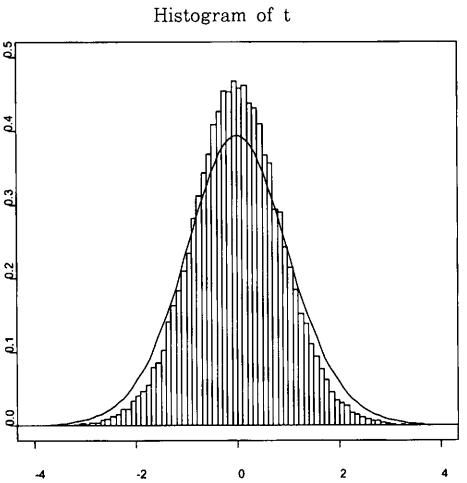

図9　2群間に0.3の相関がある場合のtのヒストグラム ($n_1 = 10$, $n_2 = 10$)

注

1) t分布は、黒生ビール(あるいはギネスブック)で有名なギネス社の技師ゴセットによって導かれた(1908年)。ギネス社は社員が研究発表することを禁じていたため、ゴセットは論文をスチューデントの筆名で発表し、そのため、t分布はスチューデントのt分布とよばれることがある。当初ゴセットはこの統計量をzと記していたが、後に、標準正規分布に従う統計量をzで表す慣習が定着したため、tと記されるようになったという(Salsburg, 2001[7])。

2）Boneau (1960)[3]では、非正規の分布として指数分布や一様分布などの確率分布を用いているが、Sawilowsky & Blair (1992)[4]では、現実の心理学データではそのような分布は稀であるとして、8種類の特異な分布を現実世界から選んでシミュレーションをおこなっている。

引用・参考文献

[1] 間瀬茂・神保雅一・鎌倉稔成・金藤浩司 2004 工学のためのデータサイエンス入門 数理工学社

[2] 芝祐順・渡部洋・石塚智一 1984 統計用語辞典 新曜社

[3] Boneau, C.A. 1960 The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, 57(1), pp.49-64.

[4] Sawilowsky, S.S. & Blair, R.C. 1992 A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin*, 111(2), pp.352-360.

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

[6] 永田靖 1996 統計的方法の仕組み−正しく理解するための30の急所− 日科技連

[7] Salsburg, D. 2001 *The lady tasting tea : How statistics revolutionized science in the twentieth century.* W.H.Freeman.