独立 2 群間の平均値の比較の際に生じる多重性の問題について

1. 等分散検定との連続によって生じる多重性の問題

独立 2 群間で平均値の比較を行う場合、一般的な統計学の解説書では、まず、分散比 F を用いた等分散検定（以下、「F検定」と呼ぶ）を行って比較する 2 群の分散が等しいかどうかを判断し、分散が等しい場合にはステューデント（Student）の t 検定を、分散が等しくない場合にはウェルチ（Welch または Aspin-Welch）の t 検定を行うと記載している。一部の書籍では、不等分散の場合、両群のデータ数が等しければウェルチの t 検定を、データ数が等しくなければコクラン・コックス（Cochran-Cox）の t 検定を行うという記載もあるようだが、コクラン・コックス法は検出力が低いので使用を避けた方がよいとされ（小林、2010）、ウェルチ法のみを記載しているものがほとんどである。

ところが、最近、F検定に続けて t 検定を行うといった手順は、検定の繰り返しであるから、多重性の問題に抵触している可能性があるという指摘が多くみられるようになってきた。

多重性の問題を説明する際によく引用される足立（1998）は、著書の中で、「検定の繰り返しが許容される（せざるをえない）局面は、きわめて例外的で限定される。」としながらも、その「例外的ケース」として、「t 検定の前提となる両群の等分散性の確認としての F検定」を挙げている。「t検定と F検定とでは検定の指向するところがまったく異なる」こと、つまり標的とするところが、平均値の差と分散の差という異なるものなので許容されるという考え方である。

一方、多くの書籍では、多重性の問題に触れていない。永田（1992）は、F検定を「予備検定」と位置付け、第 2 種の過誤の確率 β を小さくするために「F検定の有意水準を通常の値 0.05 より大きくしておく必要がある」という理由から、「F検定は有意水準 20%で行う」としているが、第 1 種の過誤については触れていない。この「F検定の有意水準を通常の 4 倍に設定する」手法は、森・吉田（1990）、Altman（1999、原著は 1991）の書籍でも紹介されている。この手法に対して、千野（2010）は、F検定は「予備検定」ではなく、t 検定と合わせて「逐次検定」と考えるべきであるから、有意水準 20%の F検定と有意水準 5%の t検定をあわせた検定操作では、全体的危険率がおよそ 24%に増大することを示し、多重性の問題があることを指摘している。また、対馬（2007）は、多重性の問題を取り上げた中で、「定説ではない」と断った上でこの手法を紹介しているが、多重性の問題の回避策としてはとらえていないようである。

F検定と t検定の併用が多重性の問題を生じる可能性があることを記載している書籍は、ほとんど見られないが、後述するように Web 上では、複数のサイトで多重性の問題が取り上げられている。一方、t検定の使いわけそのものに対する評価が、書籍や Web で多数みられるようになり、等分散性の評価を行わずにいずれかの t検定のみで検定を行うことが検討されている。

2. ステューデント法を使う

永田（1992、1996）や杉山・藤越（2009）は、2 群のデータ数が大きく異ならなければ、ステューデント法は分散の違いに対して頑健（robust）であり、有意水準が設定した α から大きくずれないことが知られているので、データ数をできるだけ揃えるようにすれば、等分散性の有無に関係なくステューデント法を使っても問題がないと述べている。つまり、群間でデータ数をできるだけ等しくすることにより、ステューデント法だけで検定が可能というわけである。

永田（1996）によると、ステューデント法とウェルチ法は、いずれも検定量 t を求める理論は同一で、ステューデント法が等分散を前提とすることによって計算を簡略化しているだけの違いであるから、ウェルチ法のみの適用で十分に思えるという。しかし、ウェルチ法は、「近似的な検定であり本当の有意水準が設定の値 α から少々ずれている可能性がある」、自由度の計算が複雑で面倒である、等分散の場合はステューデント法に比べて「検出力の低下を招く」という問題点があり、ウェルチ法を紹介していない書籍も多いことにも触れ、結局、各群のデータ数をできるだけ揃えてステューデント法を用いるのが良いとしている。

北里大学獣医学部　高岸

では、どのくらいデータ数が異なれば、ステューデント法に問題が出るのだろうか。Altman（1999）は、「2 群の分散にどれくらいの違いがあれば *t* 検定は使えないと明確にいうことはできない」としてウェルチ法との併用を勧めているが、永田（1996）は、「2 つのサンプルサイズの比が 2 倍以上異なり、2 つの（標本）分散の比が 2 倍以上異なれば、*t* 検定とウェルチ検定の結果がかなり異なる」ので、このような場合にはウェルチ法を用いるべきだとしている。また、秋山（2012）は、データ数が 2 群で大きく異なる場合（大ざっぱな目安として 1.5 倍以上程度）にはステューデントの *t* 検定は使えないと述べている。

実用面からの判断としてステューデント法を汎用してよいという意見もある。清水（2004）は、2 群の比較において解析方法を正規性や等分散性によって「杓子定規に」選択している論文を見かけることはあまりなく、両者のデータ数が 10 倍以上違う場合、外れ値や極端な分布のひずみがある場合といった特殊な例を除けば、ステューデント法で検定してよいという。ステューデント法とウェルチ法は、比較する群間でデータ数が極端に異ならなければ実際上大きな違いはないので、「連続変数の独立 2 群を比較する場合には、正規性と等分散性に言及せずにステューデントの *t* 検定を使っても、クレームがつくことはほとんどない」そうである。

3. ウェルチ法を使う

一方、ステューデント法とウェルチ法の検定結果に大きな差異のないことから、等分散性を検討せずにウェルチ法を使うべきであるという考え方も広まりつつある。今のところ、2 群の平均値の比較方法としてウェルチ法のみを紹介している書籍は見当たらないが、青木（群馬大）、千野（愛知学院大）、奥村（三重大）などは、自身の Web 上で、様々なシミュレーションの結果に基づき、等分散検定を行わずウェルチ法のみを使用した場合、有意水準が設定した α に最も近くなることから、ウェルチ法の単独使用を推奨している。筆者は、統計学者ではないので、数学的証明の正しさについては評価できないが、小林（2010）も、薬学分野では、等分散性を検討することなくウェルチ法を使用する傾向にあり、ステューデント法の使用例が減少の傾向にあることから、「2 群間の手法では、この手法（筆者注：ウェルチ法）で分析するのが最良である」と紹介している。ステューデント法と比較しても大きな検出力の差はないと言われているのがその理由である。

鶴田（2013）も、ウェルチ法は等分散の場合でも問題なく適用できるが、ステューデント法は不等分散の場合に適用すると第 1 種の過誤を生じる危険性があるという理由から、前提条件によって明らかに等分散であるとわかっている場合を除き、主としてウェルチ法を使うことを勧めている。

4. 統計解析ソフトでは

統計ソフト SPSS では、独立したサンプルの *t* 検定を行うと、Levene の等分散検定、ステューデントの *t* 検定およびウェルチの *t* 検定の 3 つの結果が表示される。ユーザは、等分散検定の有意確率が 0.05 以上かどうかを見て、どちらの *t* 検定の結果を採用するかを判断する（対馬、2007）。SAS も同様に、2 つの *t* 検定の結果と等分散検定の結果をまとめて表示し、ユーザに判断させているようである。

SAS が発売しているもう一つの統計解析ソフト JMP でも、分散が等しいと仮定した場合と等しくないと仮定した場合の両方の *t* 検定値が表示でき、表示の上での優先順位はない。また、同じ解析グループに、等分散性の検定メニューもあり、見かけ上は、SAS と同様の扱いをしている。ところが、JMP の公式解説書（Sall et al., 2004）の記載を見ると、標本サイズが非常に小さい（各群のデータ数が 3 以下）場合や、分散が等しいと仮定する妥当な理由がある場合を除き、ほとんどの状況で分散が等しくないと仮定した *t* 検定の使用が適しているとしている。

また、最近、広く用いられるようになってきた統計計算のための言語・環境である R では、t.test という関数で *t* 検定を行うことができるが、デフォルトの設定はウェルチ法になっている。ステューデント法で検定を行うためには、オプションで等分散性を明示しなければならない。

4. まとめ

生物学、医学、農学といった分野の調査・研究では、比較する群が 2 つしかないということは少なく、3 群以上の多群を設定する場合が多い。最近では、多重性の問題を回避するため、3 群以上の多群比較において *t* 検定

北里大学獣医学部　高岸

を繰り返し用いてはならないという考え方が浸透したため、2 群間の平均値を比較する *t* 検定を行うことが少なくなっている。しかしながら、2 群間での平均値の比較に、*t* 検定に代わる検定法は存在しない。まだまだ、明確な解答が得られるところまで来ていないようには思うが､少なくとも､多重性の問題に関する認識を持つことは、間違いなく必要であろう。

現代では、統計処理はコンピュータ処理が普通である。そうすると、ウェルチ法の計算が面倒だから云々といった考え方は除外されるだろう。仮説検定の考え方に従うと、等分散検定で帰無仮説が採択されても積極的に等分散性が保証されたわけではないから、ステューデント法を採用してよいのか疑問だという意見もある。結果的には、正規性や等分散性といった前提条件を考慮せずに、無条件でウェルチ法を採用するのが理にかなっているという方向性になるのだろうか。

【文献】

秋山徹監修（2012）バイオ実験に絶対使える統計の基本 Q&A　羊土社　p117

足立堅一（1998）　らくらく生物統計学　中山書店　p120

小林克己（2010）　毒性試験に用いる統計解析法の動向 2010　薬事日報社　p 48-49

清水信博（2004）　もう悩まない！論文が書ける統計　オーエムエス出版　p 56-57

杉山高一・藤越康祝（2009）　統計データ解析入門　みみずく舎　p.104

千野直仁（2010）　統計的独立性とその周辺（1）　愛知学院大心身科学部紀要、6:119-128.

対馬栄輝（2007）　SPSS で学ぶ医療系データ解析　東京図書　p75

鶴田陽和（2013）すべての医療系学生・研究者に贈る独習統計学 24 講　－医療データの見方・使い方－　朝倉書店 p176-177.

永田靖（1992）　入門統計解析法　日科技連　p100

永田靖（1996）　統計的方法のしくみ - 正しく理解するための 30 の急所－　日科技連　p185

森敏昭・吉田寿夫（1990）　心理学のためのデータ解析テクニカルブック　p63

Altman,DG （木船義久・佐久間昭訳）　（1999）　医学研究における実用統計学　サイエンティスト社　p167
(原著は、"Practical statistics for medical research" Chapman & Hall (1991))

Motulsky, H （津崎晃一訳）(2011) 数学いらずの医科統計学（第 2 版）　メディカル・サイエンス・インターナショナル　p229-230　（原著は、"Intuitive Biostatistics：A Nonmathematical Guide to Statistical Thinking.　2nd ed. Oxford Univ. Pr. (2010))

Sall, J et al. （2004）JMP を用いた統計およびデータ分析入門（第 3 版）　SAS Institute Inc.　p163-165.

【Web】

群馬大・青木　http://aoki2.si.gunma-u.ac.jp/lecture/BF/index.html　二群の平均値（代表値）の差を検定するとき

三重大・奥村　http://oku.edu.mie-u.ac.jp/~okumura/blog/node/2262　2 段階 *t* 検定の是非

愛知学院大・千野　http://www.agu.ac.jp/~chino/welcome_news/contents.html#test-for-means　平均の差の検定における 2 種類の統計量の独立性と全体的危険率

2013 年 2 月 22 日　初稿
2014 年 4 月 15 日　加筆修正

北里大学獣医学部　高岸