ライティング徒然草

エグゼクティブ・アドバイザー　林 健一

第 12 回 消えていった仮説検定 (2)

前回は，承認申請資料から消えていった仮説検定として，被験者背景の群間比較を取り上げ，この比較が実施されなくなった背景を説明した。今回は，「比較自体は現在でも実施されているが，他の統計手法に取って代わられた仮説検定」として，χ^2（カイ 2 乗）検定と Scheffe（シェッフェ）の検定を取り上げる。

まず，χ^2 検定は 2 値データの群間比較に用いられた手法である。たとえば，ある症状が改善したか否かを被験者ごとに判定した後，その症状が改善した被験者の割合を被験薬群と対照薬群とで比較するといった場合に汎用されたのがこの手法である。しかし，この検定を実施すると「Yates（イエーツ）の補正をするか否か」を判断する必要が生じる。やや乱暴ではあるが，この判断の何が問題なのかを一言で説明すると，「補正をしないと有意差が出やすくなり，補正をすると有意差が出にくくなる」ということになる。そして，補正の是非に関しては統計家の間でも見解が分かれていた。

幸いなことに，現在ではこの問題を考える必要がない。なぜなら，2 群の割合に差があるかどうかを評価する際には，Fisher の exact test（直接確率計算法，正確検定などと呼ばれるが，単に直接法と呼ぶことも多い）を用いればよいからである。この手法は，その名の通り，帰無仮説下で「試験で得られた分布」と「より極端な分布」が得られる確率を分布ごとに正確に計算する。そして，それらの確率の合計を P 値として表示する。このため，何かを補正する必要はないのである。

Fisher の直接法は決して新しい統計手法ではない。にもかかわらず，χ^2 検定が使われたのは，当時（おおざっぱに言って 1980 年代）のコンピュータの演算処理速度が悲惨なほど遅かったためである。先に記載したように，Fisher の直接法では個々の分布が得られる確率を計算するため，サンプルサイズが大きくなると，計算の必要な分布の数が飛躍的に増加する。しかも，当時は被験者背景の群間比較を実施しており，1 試験で割合の差の検定を実施する回数は現在よりもはるかに多かった。このため，Fisher の直接法を用いると，すべての解析を終了するまでにはそれなりの時間が必要となり，近似法である χ^2 検定を用いざるを得なかったのである。

現在では，モバイルノートパソコンでさえ Fisher の直接法を瞬時に実行する。こうなると，もはや χ^2 検定の出番はない。余談であるが，当時はパソコンのワープロソフトを使って文書を作

成する人は少数派であった。理由は漢字の変換に時間がかかったためで，かなやローマ字を入力した後に変換キーを押すと，一拍おいて変換候補が表示されるというまだるっこしさがあった。このため，大量の文字を打つ必要がある場合には，最初からワープロ専用機を使って文書を作成するか，テキストエディター（Windowsの「メモ帳」のような文字入力ソフト）で文章を完成した後にデータをワープロソフトに流し込んでレイアウトを整えたのである。

閑話休題。今度はScheffeの検定である。この検定は，「被験薬の高用量群・中用量群・低用量群＋プラセボ群」といった多群試験の連続量の群間比較に用いられた手法である。こうした比較に用いる手法には，Scheffe 以外にも Tukey，Dunnett，Williams の検定といったものがあった。このうち，Tukey 法は「高用量群対プラセボ群」「中用量群対プラセボ群」「低用量群対プラセボ群」「高用量群対中用量群」…といった総当たり方式の比較に用いる手法である。次に，Dunnett法は対照とする群とそれ以外の群との比較に用いる手法で，上記の4群比較試験であれば，プラセボ群と被験薬の各用量群とを比較する（用量群間は比較しない）。最後の Williams 法は，「用量が増すとともに評価項目Xの平均値も上昇（または低下）する」という単調的な関係を設定したうえで，対照群とそれ以外の群とを下降手順で比較する手法である。すなわち，まず高用量群とプラセボ群を比較し，有意差があれば，次に中用量群とプラセボ群を比較するといった手順で，有意差がなくなるまで比較するのがWilliams法である。

これに対して，Scheffe 法は考え得るすべて対比の比較に用いる手法である。具体的には，Tukey 法と同様に総当たり方式で群間を比較するだけでなく，「（高用量群＋プラセボ群）対（中用量群＋低用量群）」といった比較も実施する。しかし，ほとんどの場合，臨床試験でこうした比較をすることに意味はない。すなわち，臨床的に意味のない比較も実施してしまうのが第1の問題点で，その結果，Scheffe法で実施する比較の数はTukey法よりも多くなる。

次に，比較の数が増えると，統計学的な調整はより厳しいものになる。ここに記載した多群比較の手法は，いずれも「1回あたりの比較の第1種の過誤（本当は差がないのに，差があると結論する誤り）を調整することによって，試験全体の第1種の過誤が生じる確率を適切な水準に保つ」というコンセプトで設計されている。具体的な調整方法は手法によって異なるが，大まかには「比較の数が増えるほど，個々の比較では統計学的有意差が得られにくくなるように調整する」と考えていただいてよい。このため，同一データにScheffe法とTukey法を用いると，関心のある群間比較では Scheffe 法のほうが有意になりにくくなる。これが第2の問題点である。

臨床的に意味のない比較を実施する結果，統計学的有意差が得られにくくなる。なぜ，製薬企業はこんな統計手法を選んだのであろうか。実は，ここには悲しい事情があった。それは，当時の Tukey 法や Dunnett 法は各群の被験者数が等しい場合にのみ実施可能であったという計算上の制約である。しかし，第Ⅰ相試験といった期間の短い試験を除いて，臨床試験の被験者数がすべての群で揃うことは，まずない。このため，企業は試験の計画時点で，各群の被

験者数が異なっていても実施可能な Scheffe 法を（泣く泣く）選択せざるを得なかったのである。

その後，各群の被験者数が異なっても Tukey 法や Dunnett 法を実施することが可能となった。理由は精度の高い近似法が考案されたからで，統計解析ソフト SAS で近似計算を実行することが可能になった途端，Scheffe 法は姿を消した。ただし，毒性試験の解析では，しばらくの間 Scheffe 法が残存した。リスクを探索する毒性試験でこそ真っ先に葬り去るべきであったにもかかわらず，この手法がしぶとく生き延びたのは非臨床試験用の統計解析ソフトの対応が遅れたためである。

さらにときは流れ，現在では Scheffe 法はおろか，Tukey や Dunnett，Williams 法を目にすることもほとんどなくなった。それは，製薬企業が開発する品目に変化が生じたためでもあり，多重性の調整に関する基本的な考えが変わったためでもある。

けれども，これは別の物語。いつかまた，別のときに話すことにしよう。

あの時の歌は聴こえない
人の姿も変わったよ
時は流れた

（山上路夫作詞「学生街の喫茶店」）