総説

毒性試験における統計学利用上の注意点

浜田 知久馬

要約—統計パッケージの普及に伴い、毒性試験で用いられる統計手法も高度化しているが、我が国では、生物統計学の専門家が不足しているため、毒性試験の現状では統計学の利用について混乱が起きている。本稿では、毒性試験において統計学を誤用しないために注意すべき点、特に統計パッケージの利用において注意すべき点、特に検定と多重比較の意義、統計パッケージと決定樹の利用の際の問題点、論文における統計手法の記載の注意点を総論的に述べる。

はじめに

毒性試験を企画、実施するに当たって、統計学は実験の計画と結果の評価の2つの側面で、重要な役割を担っている。近年、特にICH（International Conference on Harmonization）の影響の現実化に伴い、統計学の重要性が離れてい強調されている。本年6月に開催された第25回日本トキシコロジー学会学術年会で、初めて統計のセッションが設けられたのも、このような流れを象徴する出来事といえる。しかしながら、統計学の利用について毒性試験の現場で混乱が起きているのを事実であると思う。その大きな背景には、我が国では生物統計学の普及が不十分なことがあげられる。欧米の主要な研究機関では生物統計家が常駐し、実験計画と統計解析の双方で支援する体制がとられているが、現在までのところ我が国では、施設内専門部にについては、専門教育を受けた生物統計家を配属する体制はない。このため実験者自らが統計解析を実施し評価することになるが、身近に統計学について相談できる専門家がいないため、統計パッケージの出力する結果を信頼にあたり、統計学に関する誤解や短縮に振り回される傾向がみられる。

長期的な展望を立った場合、根本的な解決策は、生物統計家と実験家が共同研究できるような体制を確立することであるが、この点については別個に議論、本稿では、毒性試験において統計学を誤用しないために注意すべき点、特に検定と多重比較の意義と問題点を総論的に述べる。

検定で有意になることの意味

多くの実験者が、“統計解析=検定”と考えているのではなくかかるとされるほど、医薬系のジャーナルでは検定が頻繁に用いられている。この最大の理由は、検定では有意差があるとも、有意の二つが一になり、結論が明確であるためである。検定に限らず、このような二つが一の判定をする場合、2種類の誤りが生じる可能性がある。

神様だけが知っている真実があるとして、真実は差なし・ありの2通り、検定の結果も有意差なし・ありの2通りである。本当は差がないときに検定の結果も有意でなければ、それは正しい判定、逆に本当は差があるときに有意でなければ、それは正しい判定、逆に本当は差があるのに有意でなければ、それは正しい判定、逆に本当は差があるのに有意差がつかなかったものである。統計の専門用語では、この2種類の誤りをそれぞれ第1種の過誤（αエラー）、第2種の過誤（βエラー）と呼ぶ。判定方式としては、なるべく過ちをおかしえないのが当然望まれ、αエラーとβエラーの両方が小さい判定方式を見つけたいわけだが、残念ながらこの2種類のエラーは同時に小さくできない。

わかるやすい例として裁判を考えてみよう。人間の裁判のそれはあるから、残念ながら本件の裁判が起こる可能性は0
総説

ではない。誤審には無実のを有罪と判定（αエラーに相当）する場合と、有罪のを見逃す（βエラーに相当）2種類が考えられる。この2種類のエラーを小さくするにはどうしたらよさそうだか？αエラーを小さくするためには、なるべく有罪と判断しないようにすればよい。極論を言えば、性善説の立場をとって、すべてのを無罪と判定すれば、βエラーがおさまることはない。しかしこれは本当は有罪の人もすべて無罪と判定され、見逃すβエラーが100%になる。逆にβエラーを小さくするには、見逃しを行うために厳しく裁判して有罪を増やすべき。しかし、こうするとき、無実の人が有罪になるalphaエラーが大きくなる。さて、それではどう判定したらよさそうだか。ここで、αエラーとβエラーのどちらが重大な誤りかを考えると、実在の有犯、特に死刑にでもしたら、取り返しがつかないことになる。司法の世界では“疑わしさは論ぜず”という言葉があるくらいである。検定でも基本的には同様で第1種の過誤を重要視して、αエラーが有罪確率（通常は5%に設定される）以下になるような判定方式の中で、なるべくβエラーが小さいものを採るのが原則である。αエラーを5%に設定すれば、本当は差がないのに偶然によって有罪になるのを、20回に1回以下に抑えることができる。

多くの種類の検定があるが、その結果は後方のによって統計的に表現することができる。p値は、本当に差がない（帰無仮説が正しけ）ときに、偶然によって差が生じる確率を表している。したがってこの確率が小さいときは、偶然ではない、と主張するかどうかの正当性がある。p値が小さい場合は、検定の精度が十分でないため、βエラーによって、本来ある差が見逃された可能性が高い。Nを小さくして統計的な精度を上げれば有意差が出てくるが、誤差が出て、有罪と主張しやすい、という矛盾が起こる。このため生物学的な検定法において検定でなく信頼度に基づく同定法検証が求められている。

毒試験における検定の問題点をまとめると。1）検定の結果は二乗択一であるが、複数の項目についてある程度、探索的に評価を行なう一括検査試験の目的とは若干の含蓋があり、これを実行するためには、p値などの他の統計学的な検定の利用、および生物学的考察が必要である。2）検定で、αエラーは常に保証されるが、βエラーの大きさはNに依存している。有意差がなくても、Nが小さいときは、変化が起きても、この差を保証することはできない。
総説

多重性の問題と多重比較

1つのデータに対して、1回のみ検定を行うのであれば、検定の論理は合理性を持っているが、複数の仮説を同時に検定する場合、個々の検定の有意水準をαにしても、全体としてどれか1つ以上の仮説を誤って有意とする確率は、αよりかなり大きくなってしまう。このような多重性の問題に対処するため、複数の検定の結果全体を1つの手法として考えて、全体での第1種の過誤がαに抑えられるようにしたのが多重比較の手法である。毒性試験では、対照群と複数の用量群の比較のための多重比較法として、Dunnett法が頻用されている。ただし、毒性試験で評価する仮説の構造は、より複雑である。有害事象のスクリーニングを行なうという目的のため、反復扱い毒性試験では、生化学・血液検査値、臓器重量などの数十を越える多くの項目について評価を行う。個々の項目について、Dunnett法で検定を行えば、それぞれの項目について、対照群との比較で、誤って有意差が出る確率は有意水準αに抑えられることができるが、全項目では誤って有意差が出る確率はαを越えてしまう。例えば10個の独立な項目を有意水準0.05で検定したとすると、1つ以上の項目で誤って有意差が出る確率は約0.40となる。0.05よりかなり大きくなってしまう。このような多くの項目を考慮した多重比較の方法（例えばBonferroniの不等式の利用）も考えられるが、毒性試験の現場では使われていない。その理由の一つは、何十項目もの多重性を調整すると、それぞれの項目に対する有意水準（α）がかなり厳しくなるため、その見返りとして見逃しのβエラーが大きくなるためである。毒性試験では、多重性の調整は必要ではなく、1検定の繰り返しで解析すべきであると極論する統計学者もいる。彼らの論拠は、毒性試験ではαエラーよりも、変化を逃すβエラーの方が重大な過ちであり、多重性の調整を行なうことによって、βエラーを増大すべきではないということである。このように毒性試験で、多重性の問題をどのように考えるべきかについて、統計学者の間でも一致しておらず、この点が混乱をもたらす大きな原因となっている。さて、それでは、毒性試験の多重性の問題について、どう考えればよいのだろうか？

αエラーとβエラーのバランス

検定のような2者択一の判定をする場合、αエラーとβエラーを同時に小さくはできない。したがって実的には、この2つのエラーのバランスを考える必要がある。バランスを測る上での基準は、あるべく、実験者の判断に近くなるようにすることであろう。検定の結果を最終的に評価するのではなく実験者であり、実験者が直感的に納得できるバランスが望ましい。毒性学の判定と統計的な検定の結果の一致性を評価した結果では、Dunnett検定の有意水準は両側5％、用量相関性的検定は両側1％程度が、毒性学の判断に近くなることが示された（Hamada（1997））。このような研究の結果が、バランスを考慮する上での目安になるだろう。

統計的手段の性質を理解しp値の評価

有意水準について、比較あたりで制御するか、多群全体で考えるか、多項目性まで考慮して考えるか、ポリシーの問題であり、理論的にはどれがよいか決められる問題ではない。重要なことは、手法の選択に対する主張を持ち、特にαエラー、βエラーの大きさをの性能を理解した上で、評価に用いることである。また単に有意差なし・ありだけではなく、p値に着目するのも重要である。例えばp<0.0001であれば、多重性的考慮の無有にかかわらず、偶然を越えた変化といえるが、p=0.05перед後は、微妙な結果であり、仮に有意であったとしても、αエラーの可能性は数％は残ることになる。毒性試験では多項目に渡る結果を総合的に評価して、最終的にどの用量から毒性が生じているかを判断する。したがって、各個別の検定の結果を、厳密に二択一の捉えるのは必ずしも合理的とはいえないと。

統計パッケージの利用

15年程前は、統計解析といえば、ほとんどのt検定でさえ使われていたが、最近では多様な統計手法が用いられるようになっている（浜田他，1998）。この大きな要因は統計パッケージの普及である。実験者自身で
総説

複雑な計算が必要な統計手法を実行できるようになり、統計学の恩恵を受けることができたのもは望ましいことではあるが、その無邪気な使用は多くの誤用を生むことになった（Finney, 1998）。統計パッケージを利用する上では、次の点をよく認識する必要がある。

1. 統計パッケージの出力する結果は、いつも正しいとは限らない。
   統計パッケージといっても、人間が作り出したものであり、その結果が完全に正しいとは保証できない。特に欠測が多数あるような場合、観測値の数が少ないのに多くの説明変数を用いて多変量解析を行う場合などは、しばしば計算結果が不安定になる。また計算の結果自体は正しかったとしても、それぞれの統計手法の前提条件が満たされてなければ、それはやはり統計手法の誤用である。常識に反する結果や、図示した場合の印象と著しく異なった結果が生じた場合は、前提条件を含めて、統計パッケージの出力する結果を疑ってみる必要がある。

2. 統計パッケージ間で出力する結果は、微妙に異なる。
   例えば、基本的なノンパラメトリックな検定手法であるWilcoxon検定を例にとってみるとても、ソフトウェア間で計算結果は微妙に異なる。この理由は、連続修正の有無、同順位データの処理、検定統計量の分布を正規分布で近似するか並べ替え分布を用いるかなどの細部の点について、様々な流儀が存在するためである。統計手法によっては、どの流儀で計算すべきか、定説がないものもある。例えば2×2の分割表のFisherの正確検定の偏側確率の計算手法については、統計学者の間でも意見が大きく異なる。したがって統計パッケージを選ぶ際には、正確にどのような流儀で計算しているか、マニュアルに記述されているか確認しておく必要がある。また以下の研究機関の中では、統計パッケージは必ずしも同一のものを選んで標準化しておくことが望まれる。

3. 汎用的かつ信頼性の高いものを選ぶ。
   多くの統計パッケージが存在するが、その信頼度は質的に大きく異なっている。毒性試験も国際化時代を迎えつつあり、世界的に広く使われている信頼性の高い統計パッケージを選ぶ必要がある。

決定樹

多くの統計手法の中から、得られた実験データに対して適切な統計手法を選択するの、なかなか困難で長い経験が必要である。統計に関する知識が十分ではない実験者が、適切な統計手法を選択するためにしたのが決定樹である。判別も少決定樹であるが、それなりの必要性があり、また大きな役割を果たしていると思う。例えば、計量データがある分野では、外れ値について検討し、次に分布の型を検討し、それから等分散性を検討するといったように、細かく分岐して入って、最終的な統計手法をたどり着く（吉村、1987）。解析の目的やデータの型・前提条件に応じて、枝別れるする決定樹は、データ解析の理論的な流れを示す上で有効である。ところが実際には、理論的な流れが無視されて、最終的に選ばれた統計手法で、有意になったかどうかのみが、強調されすぎている傾向がある。極論を言えば、最終的に選ばれた手法が何であるかはすら注意が払われず、有意であるかの一点に関心が集中する。しかしながら、特に毒性試験の評価は探究的な側面が強く、解析のプロセスを追うことは必要なことである。最終的には有意差がつかなくても、外れ値の存在・分布の歪み・不等分散性は、毒性の兆候を示している可能性がある。

機械的に決定樹を利用するのでなく、それぞれの分岐（予備解析）の生物学的意味をよく考えることが必要である。また選ばれた統計手法が、どのような特徴があり、性能がどの程度異なるかを認識しておくことも重要である。統計学自身も常に進歩を続けており、決定樹も歴史の流れとともに変化するものであるから、常に最新の情報を入手するように努める必要がある（Hamada, 1998）。

論文における統計手法と結果の記載

論文における“Statistical Analysis”のセクションに
総説

毒性試験における統計学利用上の注意点

おける記述の原則は、論文の読者が、類似データに対して同じ解析を行うのに十分な情報を提供することである。客観的な科学論文である以上、第三者による評価を含めた実験の追試を可能にする必要がある。結果については、“Statistical Analysis”の記述と矛盾がなく、しかも対応を明らかである必要がある。決定樹を用いる場合、毒性試験の全ての項目について、その道筋を示すことは無理としても、考察・結論に関連するような重要な項目については、単に最終的な有意差の有無だけではなく、必要に応じて等分散性の有無、外れ値の存在等を記述する必要がある。

以下いくつかの注意点を箇条書きで示す。
1. 図・表におけるデータの要約方法を明記する。
   特にあくの表現では、mean±SD なのか mean±SE なのか、それ以外なのかを記す。
2. 動物数 N や自由度、解析の単位を明示する。
3. 高次の分散分析を用いる場合は、モデルを明らかにする。
4. 解析目的と項目ごとに適用した統計手法を明記する。
5. 検定を行なう場合、事前に決めた有意水準、片側か両側かを明らかにする。
6. 計算に利用した統計パッケージを明記する。
7. 検定では原則的に p 値を示し、必要に応じて信頼区間に示す。

おわりに

統計学の形式的な利用はたいへん危険である。毒性試験の実験計画、データ解析については、事前に統計の専門家によく相談することが重要である。

参考文献

浜田知久馬、小野英樹、板東正博、吉野慶（1998）：薬効・薬理試験の統計的評価とその問題点、第66回日本統計学会講演報告集、1–2。
吉村功編著（1987）：毒性・薬効データの統計解析、サイエンティスト社、東京