原 著

全日本鍼灸学会誌における統計処理の問題点

(社)全日本鍼灸学会

研究委員会・情報・評価班

七 堂 利 幸

要旨 全日誌発刊(1982年32巻1号)から、1991年現在(41巻2号)までを調査したところ、統計処理を行っている169報告中、164報告(97%)は何等かの統計処理の問題があった。その中で基礎の105報告中は99報告(94%)、臨床の64報告中62報告(97%)は、両者の区別なく同程度に統計誤用があった。

最も多い問題点は多重性の問題で、問題のあった全報告(164)のうち45%(73)も統計誤用があった。次が、検定法、解析方法無記載が19%(31/164)、群内検定で結論を出してしまっている報告13%(22/164)、有効率の提示のみで不確実性指標がない報告14%(23/164)、そして無作為化の保証が無い報告12%(19/164)であった。全体では18項目のチェックポイントがあるが、その他は統計誤用は各々10%未満であった。

ただし、これは各報告に複数の問題があるので、全体の集計は100%を越える。

比較的结果に大きな影響を及ぼさないと考えて、検定法・解析方法無記載19%と、有効率の提示のみで推定がないもの14%を除いても、統計処理を行っている雑誌全体の62%に統計誤用があることになる。

論文著者の側からみると、11人の著者で全体の30%(49報告)の問題を起こしている。つまり同一人が3~7報告、繰り返し誤りや問題を起こしている。この中で多重性の問題があるのは91%、次が群間で比較すべきを群内検定しているものが55%と、目だっている。

I はじめに

医学論文に統計処理は不可欠になってきている。しかし、統計誤用は一流の学会誌にも頻繁に起こり¹⁾、不注意からくる簡単なミスから結論を大きく変える誤りまで、その実状、警告、そして対策が報告

されている^{2~5)}。

全日本鍼灸学会誌(以下、全日誌と略す)に対する統計処理誤用の部分的な指摘は、過去にあったが⁶⁾、全般的な調査は行われていなかった。そこで、全日誌の統計処理の問題点が、どのくらいの割合で発生しているか調査し、今後の対策の基礎資料としたい。

II 方 法

全日誌発刊(1982年32巻1号)から、1991年現在(41巻2号)までを対象とした。

* Toshiyuki SHICHIDO 全日本鍼灸学会研究委員会、情報・評価班

Key Words : 統計誤用、尺度合わせ、多重性、検定、推定、有意差、独立性、群間検定、群内検定、不完全データ

論文の種類で基礎と臨床の区別は、直接臨床効果を探っているものを臨床、それ以外で患者を対象としても、刺激実験を行っているものは基礎とした。

以下のような問題点(重複あり)について、その頻度を調査した。

Ⅲ 問 題 点

1. デザイン不良

試験、実験のデザインそのものが悪くて、とても出されている結論は導き出されないもの。たとえばサンプルが少なすぎるため、検出力が少なく有意差がでないもの。

実験者の盲検は考えられても、評価者の盲検は無視されている。実験者のバイアスが混入している。

他人による追試が不可能なデザインは、あるいはプロトコールが明記されていないものは、報告としては信頼性がない。

2. サンプル、評価項目の独立性無視⁵⁾

検定を行う時、個々のサンプルは独立であることが前提である。関連ある評価項目をどんどん検定しても、有意差がでやすくなるだけである。そのためには無作為化が必要になる。例えば、空腹感、摂食量、そして満腹感の得点を合計して、食欲抑制係数というものを作り、群間比較するというやり方である。前三者はお互い相関関係にあり、独立でない。自作の食欲抑制係数という、独立でないものを加算したものでサンプル数を増やし、有意差をだしやすくなるというものである。

3. 尺度合わせ^{4,9,11)}

例えば著効、有効と、やや有効、無効を一緒にして、二段階で検定すると元の四段階のまま検定するのとは比べ、有意差がでやすくなる。これは多重性の問題でもある。

4. 多重性の問題^{7~11)}

多群比較、多項目比較、多期比較、多層比較いずれも、有意差がでやすくなる。そこで多重性の考慮が必要である。特に経時的データにおいて、連続した時点の反応は通常高い系列相関があり、独立でない。特に多重t検定の論文が多い。

対策は評価方法を考え、かつ多重比較、くりか

えし測定を持った多要因分散分析、多変量分散分析、対数線形モデル等の適用を考えるべきである。

①輪切り検定

各時点毎に対照群と処理群を2標本t検定等で比べる方法。全体的なパターンを考慮していないので、部分的に有意差が出ても、現象の整合性ある説明が出来ない。

②初期値との差検定

初期値と各時点の差を求め、t検定等を繰り返すもの。

③差の差検定

初期値と各時点の差を求めこの群間を比較するもの。

④多層比較

性別、疾患別、重症度別等を事後に層別。

⑤多項目比較

血圧を最低血圧、最高血圧、その差といった具合に、分けて項目を増やし、あるいは血液検査の全項目比較検定する類のもの。

⑥多数層別検定

後知恵で細かく任意に層別し、それらについて検定を繰り返すもの。

⑦多重目標(endpoint)

多数のエンドポイントについて、検定が施される。

⑧多種検定

例えば分割表の場合、 χ^2 検定、U検定、累積 χ^2 検定、最大 χ^2 検定を行い、後で有意差のあるものを選ぶといったもの。

5. 順序尺度の検定ミス^{1,11)}

順序尺度で評価したものを、順序を無視した χ^2 検定で計算すると感度が鈍くなることがある。

6. 有意差なし同等性¹¹⁾

有意差がないことが、即両群同等といえない。症例数を増やせば有意差が出る。ネガティブトライアルの考え方でサンプル数の設計をしてからでないと、両群の同等性は証明できない。

7. 無作為化の保証なし⁵⁾

無作為化は、様々のバイアスを各群均等にばらつかせる効果がある。無作為化の保証無しに検定しても意味が無い。ただ無作為割付を行っても観測値に独立性の無い場合もある。例えば10名のサ

ンプルしかないのに重複をして、5群30名の比較検定をしようとする類である。

無作為化試験ではない、観察的研究の場合、無作為化していないのであるから結果にどう影響するかの吟味が必要。無作為化のない研究は、サンプルから考えられる母集団は何なのか明確でない。

8. 持ち越し効果, 順序効果無視

クロスオーバー法のような方法がとられた場合、前期の効果が後期の治療に影響することを持ち越し効果、前期と後期の順序を入れ換えたときその効果が変わる場合、順序効果があるという。これらを見捨てて結論を出している場合。

9. 不完全データの扱い

脱落、中途打ち切りデータ等を除いて解析したり、その記載が無かったりする場合。都合のよい結果を導くことは可能である。

10. 比較方法のミス

無作為同時対照試験でないものを、つまり、症例数が集まりにくいいため、ある長期期間内で集めた短期の試験期間の症例で、しかも時期により自然変動、自然治癒、プラセボ効果の起こり易い疾患を扱う場合、同質の群として2群の比較をしてよいものかどうか疑問である。時期により偏りが無いかどうか保証する必要がある。

歴史的対照と比較するもの、現在あるカルテを適当に分け、群間比較しようと言う時、無作為同時対照試験と同じ結論は導き出せない。

11. 群間でなく群内検定で結論

比較する対照群を設定しているにもかかわらず、結論は群内検定の結果をもって、さも群間で有意差があったという結論を出している場合。何のための対照群か不明であり、比較の意味が理解できていない。

または対照群を設定せずに、初期値との群内検定で、さも群間で比較したかのような結論を出している場合。いずれも誤りである。

12. 検定方法の疑問

各検定法にはそれを使うための、条件がある。分布の位置、形に敏感な検定法があり、ノンパラメトリック検定法も例外ではない。例えば、t検定や分散分析は、①等分散、②正規分布、③無作為化の条件を同時に満たさないと使えない。

そこで等分散、正規分布の条件がない場合、ノンパラメトリックな方法を採用してもよいが、無作為化の条件は必要である。

10や20くらいの少ないサンプル数で正規性を議論することは容易でない。等分散でも安易にt検定を使えない。

また有意差がでたが、関連は無いという場合もある。

13. 検定名の記載なし、統計処理の方法の記載なし

数多い検定法あるいは統計方法の中から、どの検定、方法を使ったか記載するのがルールである。検定法、統計処理の方法も研究方法の一部である。読む人がチェックできるよう記載が義務づけられる必要がある。また、コンピューターのプログラムパッケージ名も、同様の理由で記載を義務づける必要がある。

14. 推定等による不確実性の指標なし^{4,12,13)}

有効率はその試験、実験によって設定されたものの、だから異なった報告の有効率を比較することは容易でない。試験のデザイン、評価尺度の決め方、症例数等で変わって来る。

また、有効率だけ何%と報告しても、その不確実性が分からない。有効率は症例数、試験デザイン、評価尺度の決め方などによって変わり、その試験に限定されたものにすぎない。臨床的にこの程度以上改善が得られれば、それを有効としようという合意がないと、日常臨床のめやすにはならない。単に順序尺度の著効と有効を有効とし、それ以下の無効と比較して有効率を計算しても、一般化、普遍化できない。単に有効率を計算するだけの報告の有用性が問われる。

小症例数の場合の有意差の議論は問題である。これは有意差があるとかないとかという是か非かの議論より、信頼限界を示した方が良いという意見が多い。有意差はあってもその差が臨床的に意味を持たない程小さい場合、また有意差がなくともそこに重大な差が存在する場合、信頼限界(これも症例数に依存するが)の方が有意差検定より良いという考え方である。

15. 個々の対象に関するデータの欠落

分布の形も検討せず、平均値±標準偏差(あるいは

は標準誤差)だけで結果を表している報告がある。その前に $m \pm S.D$ か、 $m \pm S.E$ か記載のない報告がある。

平均値 $\pm$ 標準偏差(あるいは標準誤差)だけでは、個々のデータの反応が読み取れない。そのためには個々のデータをグラフや表で表す必要がある。

16. 不適切な評価項目

順序尺度や、症状に適当に点数をつけ、それを集計して解析。肝炎の研究に不適切な肝機能検査項目を採用しているというようなこと。

評価方法に信頼性、妥当性の検討が済んでいないのに、その評価方法で評価している誤り。

17. 多変量解析の乱用

主成分分析、判別分析、因子分析などを使っているが、元のデータの信頼性、妥当性があるのか、またそれが定量出来るものかどうか疑問なもの、また多変量解析が何のために使われているか、結果の解釈が明確でないもの等がある。

18. 結果の一般化の言い過ぎ

例え無作為な標本抽出でも、その特定した母集団しか一般化できない。例えば分散分析では、母数効果(fixed effect)モデル、変量効果モデル、そしてその混合効果モデルの3種のモデルが考えられ、独立変数の一般化には変量効果(random effect)モデルに従う必要がある⁵⁾。

また、一定の条件での試験ではその結果の一般化は非統計的な推論に依らざるを得ない。複数の結果でなく一回の実験で一般化することは危険である。

IV 結 果

全日誌発刊(1982年32巻1号)から、1991年現在(41巻2号)までを調査したところ、統計処理を行っている169報告中、164報告(97%)は何等かの統計処理の問題があった。その中で基礎の105報告中99報告(94%)、臨床の64報告中62報告(97%)は、両者の区別なく同程度に統計誤用があった。

最も多い問題点は多重性の問題で、問題のあった全報告(164)のうち45%(73)も統計誤用があった。次が、検定法、解析方法無記載が19%(31/164)、群内検定で結論を出してしまっている報告13%(22/164)、有効率の提示のみで不確実性指標がない報

告14%(23/164)、そして無作為化の保証が無い報告12%(19/164)であった。全体では18項目のチェックポイントがあるが、その他は統計誤用は各々10%未満であった。

ただし、これは各報告に複数の問題があるので、全体の集計は100%を越える。

比較的结果に大きな影響を及ぼさないと考えて、検定法・解析方法無記載19%と、有効率の提示のみで推定がないもの14%を除いても、統計処理を行っている雑誌全体の62%に統計誤用があることになる。

論文著者の側からみると、11人の著者で全体の30%(49報告)の問題を起こしている。つまり同一人が3~7報告、繰り返し誤りや問題を起こしている。この中で多重性の問題があるのは91%、次が群間で比較すべきを群内検定しているものが55%と、目だっている。

V 考 察

多重性の問題は1965年前後にわが国の臨床医が指摘していたという⁸⁾。New England誌では、1970年代には警告の論文が続々発表されている¹⁾。

その後、1980年Circulation誌に多重比較の問題が取り上げられ、その後日本から投稿した論文が拒否(reject)されたことがきっかけで、日本でもこの問題にたいする関心が高まったという¹¹⁾。

全日誌が創刊されたのは1982年であり、創刊以来現在まで、約3年平均でみても、平均してこの多重性のミスが現れている。急に最近増えたものではない。統計処理している報告の45%がこの誤りを犯しているということは、重大な問題である。ネガティブの報告は投稿しないという傾向を考えると、これだけの報告に言い過ぎがあるということである。

多重性の問題以外の統計誤用は、各々、全報告の20%以下であるが、これも発刊以来、平均して掲載されており、減少傾向が無い。

同一の論文著者が、最大7報告に同じ誤りを繰り返している。同一著者で、3報告以上統計誤用のあるものを集計すると、全報告の30%にもなる。もしこれら11人の著者の報告が改善すると、一挙に30%もの論文が改善する。彼らの誤りのほとんど

どは多重性の問題と、群間ですべきを群内検定で結論を出しているものである。二つとも結果に大きな影響のある統計誤用といえる。

投稿者が世界的な範囲の New England 誌では Medical Uses of Statistics という過去の統計誤用に関した論文集が、その他論文の質を高める方法の論文と共に一冊の本となり1986年出版され、この方面の教育啓蒙に一役買っている¹⁾。

全日誌の質を高めるために具体的には、研究者の教育啓蒙が必要である。全日誌反復投稿者は比較的限られた範囲の者であるし、査読時の指導、研究者のためのマニュアル作成などで、ある程度までは教育啓蒙の実効を上げ得ると考える。

引用文献

- 1) J. C. Bailar, F. Mosteller : Medical Uses of Statistics, NEJM Books, 1986.
- 2) S. A. Glantz : Biostatistics : How to detect, correct and prevent errors in the medical literature, Circulation, 61(1), 1-7, 1980.
- 3) D. G. Altman, et al : Statistical guideline for contributors to medical journals, Bri. Med. J., 286(7), 1489-1493, 1983.
- 4) 長谷川宏明ほか：「医学のあゆみ」誌“くすり欄”の臨床比較試験結果の要点一覧表，医薬ジャーナル 17(10);95～99.(1981)
- 5) 橘 敏明：統計的検定の誤用と弊害，医療図書出版，(1986)
- 6) 七堂利幸：医学統計から見た鍼灸論文の読み方<4>，鍼灸OSAKA 4(1);52～60.(1988)
- 7) C. R. Buncher, Jia-Yeong Tsay 編：医薬品開発過程における統計的方法(佐久間昭ほか訳)，MPC, 1990
- 8) 藤田利治ほか：多数反応項目およびサブグループに対する統計的検定の繰り返し適用の問題点，臨床評価 12(3);827～835.(1984)
- 9) 広津千尋：薬効検定でよく用いられる統計的方法とその問題点について(2)，臨床評価 12(3);589～610.(1984)
- 10) 佐久間昭：くたばれ t 検定，ファルマシア 24(6);568～572.(1988)
- 11) 斧田大公望：医薬統計解析マニュアル，メディカルリサーチセンター，141～142.(1988)
- 12) K. J. Rothman : A show of confidence, N Engl J, 299(24), 1362-1363, 1978.

(〒543 大阪市天王寺区上本町6-3-31-1111号)

Statistical Problems in Papers in the Journal of Japan Society of Acupuncture

Toshiyuki Shichido

Information & Evaluation Reserch Group, in Japan Society of acupuncture

ABSTRACT

In investigating all the back articles of the Japan Society of Acupuncture Journal (vol.32 no.1 1982) until the present (vol.41 no.2 1991) it was found that out of 169 articles using statistical analysis 164 of them (97%) contained some kind of problem. Of these, 99 out of 105 (94%) stimulus experimentations by animal and human, and 62 out of 64 (97%) clinical studys were found to contain misuses of statistics approximately the same percentage for each group.

A high incidence of problem was of multiplicity, with 73 out of 164 problem articles (45%) containing misuse of statistics.

Beside this, 19% (31/164) of the articles failed to list the methods of analysis or test of significance used, 13% (22/164) published their conclusions based only on comparisons within a group, 14% (23/164) published only the rate of effectiveness, without reporting the rate of uncertain results, for example, confidence interval, and 12% (19/164) failed to give assurances against randomness. Overall there are 18 checkpoints, but there is a less than 10% rate of statistical misuse in each of the others.

Because each report has multiple problem the total percentage is greater than 100%. Even excluding the 19% of the articles failing to list methods of analysis or test of significance and the 14% listing only the rates of effectiveness, neither of which ultimately affect the article results, there are still statistical misuse in 62% of the articles using statistical analysis.

Looking at the authors it can be seen that 11 writers account for 30% (49 articles) of the problem. In short, the same people are repeating the same mistakes in from 3 to 7 articles each.

Of these, 91% contain problems of multiplicity, and 55% contain comparisons conducted within a group instead of between groups as required. These are their most glaring faults.