

前臨床試験の統計学的側面

吉 村 功*

Statistical Aspects of Pre-clinical Experiments

Isao Yoshimura*

This paper surveys statistical issues appeared in pre-clinical experiments. From methodological point of view, strategies for multiplicities, estimation of effective dose, analysis of sequential data, and utilization of historical controls are discussed. From toxicological point of view, problems related to carcinogenicity, mutagenicity, teratogenicity, and behavioral toxicity are discussed. The author anticipates that much progress will be achieved if the importance of close collaboration between toxicologists and statisticians are realized.

本稿は、動物や細胞を使って毒性や薬効を調べる試験について、現在焦点となっている統計学的な課題のサーベイを試みている。統計学の方法論的な分け方からは、多重性への処理、パーセント影響量の推定、経時データの解析、背景対照の利用を取り上げ、毒性試験の分け方からは、発癌性、変異原性、生殖毒性、行動毒性を取り上げている。新たに特別な理論や手法を用意しなくても、毒性学の学者・研究者と統計家の間で十分な知識の交換と共同作業が行われれば、生じている多くの問題は解決すると予想している。

1. はじめに

本稿では、前臨床試験の統計学的側面で問題になっていることを、統計家に応用分野での問題を示すというやり方で概観する。

われわれのまわりには、薬品のように毒性を押さえつつ薬効を利用する物質と、環境汚染物質のように好まざる存在として毒性を持つ物質とがある。これらの物質について薬効や毒性を調べるには、動物や細菌を使った実験で、薬効と毒性の発現の統計学的あるいは生物学的有意性を吟味するのが普通である。

薬品等の場合には、これらの実験は、ヒトでの実験すなわち臨床試験の前に行われる。それゆえこれらはしばしば前臨床試験と総称される。環境汚染物質等の場合には、これらの実験は、疑われている毒性を確かめるために行われ、ヒトを用いることはないから、前臨床試験と呼ばれることはない。単に毒性試験と呼ばれるのが普通である。しかし実験のやり方やデータの解析法には両者に共通点が多いので、ここでは両者をあわせて議論の対象にする(白須ら[38])。

前臨床試験の統計学的側面を概観するには、二つのまとめ方がある。一つは統計的方法に応じて問題を整理するものであり、他の一つは試験法の目的に応じてそれを整理するものである。統計学者が中心になっている関連雑誌、たとえば *Biometrics* や *Statistics in Medicine*, での論文や統計学者が編集した書物、たとえば吉村[52], では前者の形態が多く、毒科学者や薬学者が中心になっている雑誌、たとえば *Mutation Research*, *Journal of National Cancer Insti-*

tute, Risk Analysis, での論文や毒科学者が編集した書物, たとえば Krewski ら [22], では後者の形態が多い。ここでは吉村・大橋 [53], Peace [32], Hothorn [17] などにならって, 両者を併用して説明を行う。なお以上に挙げた著書は, いずれも本稿で対象としている話題についての総括的な文献である。

前臨床試験の目的は, 問題の薬物に目指す薬効となんらかの毒性があるかないかを確認することと, もし毒性があるならその毒性がどの程度であるか, その毒性の発現の機序がどうかであるかを評価・確認することである。薬品の場合には, 毒性が大きければ, その段階で開発が中止される。それゆえ前臨床試験のうちの毒性を確認する部分は, しばしば安全性試験と呼ばれる。

多くの場合薬効は, その存在が過去の知見からある程度予想されていて, 注目すべき特性(endpoint)もどちらかという狭い範囲に限定されている。これに対して毒性は, 予想もしなかったところに異常現象として現れ得るので, 前臨床試験での毒性吟味では, 被験動物をあらゆる側面から予断を持たずに調べなければならない。したがってそこでの統計解析では, 毒性についての課題が, 薬効の確認についてのものより多くなる。当然, 統計学的に議論すべきところも多い。

薬品については, 多くの国で生産・使用に制約がつけられ, 規制が行われている。規制においては, 各国ごとに, 行うべき実験や調べるべき項目が規定されている。しかし, 実験の計画やデータ解析の手法等の細部は, 薬品の製造あるいは輸入の申請者にまかされているので, 実際にどのような計画で実験を行い, 得られたデータをどのような手法で解析するのが合理的かは, それぞれで独自に研究すべき課題となる。また, 実験の計画と解析の標準の国際的な調和も課題の一つになる。人種や生活習慣が関係するので, この課題は解決が容易でない。

規定されている試験は, 大きく分けると, 一般毒性試験と特殊毒性試験になる。前者は成長抑制や臓器への障害等, 古くから認識されている毒性を調べるものであり, 後者は癌を誘発する発癌性, 生殖機能を傷害する生殖毒性, 遺伝子への変異をもたらす変異原性, アレルギー反応に関係する抗原性, 麻薬などのように習慣的使用を呼ぶ依存性, 局所に特別な影響をもたらす局所刺激性等, 特に注目を浴びた事件を通じて認識された毒性を調べるものである。

前者における統計学的課題としては, 多重性による第1種の過誤の上昇や反復測定によるデータの相互相関性への対処といった, 臨床試験などと共通で, かつ従来から一般的にさまざまな論じられていて未解決なものが多いのに対し, 後者では特別な状況下での, 独特な問題への対処法の吟味・検討が問題になる(渡邊ら [46])。

2. 多重性への対処

ここで多重性として問題にしていることは, 一連の実験に対して複数の仮説を考えて統計的検定を行うことで, 第1種の過誤が公称の(nominal)有意水準を大きく上回ってしまう現象のことである。

薬効では, 多重性を適切に考慮しないと, 第1種の過誤による薬効の過大評価が起こることから, 臨床試験のデータ解析では, 一時期, 多重比較法の利用が強く求められた。この数年, 多重比較法の論文, 著書, 報告が多く出ているのはその影響であろう。こうして登場した多重比較手法を概括的に調べるには, Hochberg ら [10], 統計数理研究所研究リポート 18 [43], 同 22 [44], 吉田道弘 [54] が利用できる。この中で, Hochberg らの著書は教科書的構成で最も詳しい議論を展開しているが, 発行が1987年であるため, その後発表された議論は載っていない。

やや古い書物の多重比較法の説明には, 多群の実験での任意の群間差の検定では, 群の大き

さ(群内の動物数)が等しいときは Tukey の方法, 等しくないときには Scheffe の方法を使うべきである, という記述がある。これにもとづいて手順を定めている企業がある。しかしそれは適切でない。群の大きさが等しくないときは, Tukey の方法を修正した Tukey-Kramer の方法を使えばよい。これを示したのが Hayter [09] である。

古典的な多重比較法の欠点は, 第1種の過誤を重視して検出力を軽視したことである。多重比較法が多用されるようになって, 検出力の軽視に対する反省が生まれ, より高い検出力を得るための手法が工夫された。主たる工夫は, 多段階手順を用いることと P 値の順序統計量を用いることである。手順は複雑になるが確かに検出力は大きくなる。これらを議論しているのが, Holland [14], Hochberg [11], Hommel [15], Falk [03], Hothorn [16], 松田ら [28] である。Tukey [45] も総括的なコメント述べている。

毒性試験では, 処理群が, 用量を変えた同一薬物の投与で構成されることが多い。このときの群間比較は, 単純な多重比較ではなく, 用量と共に反応が増加することを前提にした多重比較で行う方が合理的である。毒性が無視できる程度かどうかの判定をしたいときは特にそうである。このための手法としては, Williams や Shirley が提案を行っているが, 栗木 [23], Mudholkar [31] も独自の議論をしている。しかし実際にどのようなやり方をするのが合理的かについては, まだ議論の余地が残っている。

多重比較法は一般に, 第1種の過誤の確率の最大値, すなわち FWE (FamilyWise Error Rate) を有意水準以下にすることを焦点にして作られている。そこでは検出力の確保が軽視されている。ところが毒性試験の場合には, 測定項目が非常に多い上, 一つの実験での群の数も臨床試験より多い。これらについての多重性を, 単純な多重比較法の採用で対処しようとする, 検定の検出力が非常に低くなり, 結果として毒性の隠ぺいにつながることになる。そこで毒性データの場合には, 検出力を落とさないような多重性への対処が必要になる。方針としては評価のための指標を少数に総合化することが提案されており, いろいろな試みがなされている。たとえば P-値の利用 (Selwyn [36], Wright [49]), 主成分分析等による特性値の総合 (森田 [29]), データ総合における信頼区間方式の利用 (Lehmacher [26]) などである。しかしまだ, 一般的に有効な手法が具体的定型的なものとして整理されたとは言えない状況である。

3. パーセント影響量の推定

薬物の毒性がどの程度強いかを示す指標として, かつては LD 50 (半数致死量, 50% Lethal Dose) を用いるのが常識であった。しかし近年では, ① LD 50 は, 種, 性, 齢, 薬物の摂取条件によって値が大きく変化する, ②致死率の用量反応関係は薬物によって勾配が異なるので, 単一パラメータで致死的な毒性を評価するには無理がある, ③近年の現実の使用条件では, LD 50 でとらえられる致死毒性の発現が生じるのは例外的な場合だけである, ④ LD 50 を求める実験は残酷性が強く動物愛護の観点から避けたい (Zbinden [55], Tattersall [41], Moor [30]), という理由で, 毒性が強い毒物以外では, LD 50 を高い精度で求める必要はないという考えが認められるようになった (渡邊ら [46])。ある程度の用量で実験をしたら, LD 50 の区間推定を行えば十分で, その区間を狭くするために実験をやり直す必要はないというわけである (Williams [48])。

しかし死亡に限定せずもっと一般に毒性発現の強さを考えるとき, 被験群の毒性発現率が 100 p % 増加する用量を求めることは, 相変わらず重要である。たとえばある種の腫瘍については, 発がん性の強さが LD 10 に近いのでこれを指標にできるという報告があるし, ドレイズ試験の代替法としては各種細胞への 50 % 影響量 (50% Effective Concentration, EC 50) がよいという報告 (小島ら [19]) もある。発癌性の強さについては, 一生に対応する期間において

腫瘍を50%増加させる用量(50% Tumor Dose, TD 50)で把握するという議論も見られる(Sawyerら[35])。したがって、100p%影響量を推定するための手法はまだ研究の余地がある。これについては、Hoekstra[12]が総合報告をしている。

4. 経時データの解析

毒性の発現は薬物の摂取後徐々に強まり、頂点に達した後また弱まってやがて停止する。そこで反応が体重や血液成分のように非破壊的に測定できるものである場合には、データが経時的、時系列的なものになることが多い。このようなデータを反復測定値(repeated measures),あるいは経時データ(sequential data)という。

経時データの解析が難しいのは、個体の反応に個体固有の偏りと生物体の特徴としての生理的変動の連続性とがあって、これらが誤差的変動に相関を生じさせ、しかもその変動の相関構造自体が、個体によって異なる可能性があることである。このように個体ごとにさえ異なり得る相関の構造を、少数の観測データから推定評価することは一般に極めてむずかしい。これが経時データの解析が統計学的な課題となる理由である。これについては、Crowderら[02]が近年の研究成果をまとめているが、内容は必ずしも十分でない。手法だけについて言えば、吉村ら[53]の方がよい紹介をしている。1993年7月に行なわれる日本統計学会チュートリアルセッションの報告も、やがて公開される予定である。

5. 背景対照データの利用

前臨床試験として行われる実験では、原則として対照群が用意される。実験には一般にブロック効果が存在するので、データ解析では、原則として当該実験内の対照群、すなわち現対照(current control)のデータを使って処理群の効果を見る。ところが近年では、非常に多量の動物実験が行われており、その実験での対照群のデータが蓄積されている。このデータを背景対照(historical control)データという。

背景対照データは量が多いから、ブロック効果が小さければ現対照のデータより情報が多い。そこで次のような条件があるとき、背景対照の利用が実験者から強く求められる。①イヌやサルなど大動物の実験で、群の大きさが2~4程度に小さいために検定の検出力が低いとき、背景対照を利用して検出力を大きくしようとする。②たとえばアザラシ症のように、現対照では極めて稀にしか出現しない反応が処理群に出現したとき、背景対照を利用して処理群での出現が統計的に有意であることを示そうとする。③背景対照が非常に安定でブロック効果が無視できるとき、現対照に背景対照を加えて検定の検出力を上げようとする。④多重性の悪影響で第1種の過誤としての有意差が出現したと思われるとき、背景対照を参照して統計的な有意性を否定しようとする。

このような要求に応じてTarone[40], Hoelら[13]は、背景対照の利用手法を提案している。一方Margolinら[27]は、利用の理論を提示しながら、背景対照を利用するのに最低限必要な実験上の条件を示し、背景対照の乱用を戒めている。またTamuraら[39]やKrewskiら[20]は背景対照の利用が、ある条件の下で第1種の過誤の大きな上昇をもたらることを、シミュレーションで示し、背景対照の利用を注意するよう指摘している。それでもYanagawaら[50]は、②のような場合には背景対照を利用することに意義があると主張し、保守的な(第1種の過誤を十分低く抑制する)利用ならばよいと報告している。Hayashiら[08]は③の場合に背景対照の利用が検出力を増加させることを示している。Yoshimuraら[51]は④の場合での背景対照の利用は危険であることを注意している。背景対照の利用は現実にはよくなされているが、その利用法はまだ研究の余地を残している。

6. 発癌性における問題

発癌実験のやり方は、げっ歯類を2年程度飼育して癌（悪性腫瘍，悪性新生物）の発生頻度を調べるものである。非常に大きな労力・費用がかかるので，群の大きさは1群50匹程度に限定せざるをえない。統計的な検出力を考えて群の大きさを定めることはむずかしく，動物が長期の実験に耐え得る範囲でなるべく大きい用量（Maximum Tolerated Dose: MTD）で実験を行うことで検出力を高めようとする。そこで統計学的な問題は，MTDの適切な推定法とデータの解析法にしばられる。前者については今のところ，説得性のある統計的な議論はなく，論文として報告されているのは後者についてのものが主である。

特定の型の癌（悪性腫瘍，悪性新生物）に注目しているとき，問題は生存時間解析（survival data analysis），寿命データ解析（life time analysis）と同じで，癌の発生までの時間分布が処理によってどう変わるかである（Lee [25]）。しかし，単純な生存時間解析と状況が異なるのは，単に癌を生じさせたかどうかだけでなく，それが生命を脅かすレベルのものかどうかの問題なことである（Petoら [33]）。

たとえば飼育期間中に死亡した個体が，解剖によって癌細胞が発見されてしかも他の病状も発現していたとき，この個体の寿命が癌によって縮められたものとしてよいかどうかである。これに関連した解析法についての議論は非常に沢山あるが，まだ決着がついていない。

特定の型の癌に限定せず，発癌性一般を考えると，問題はさらに複雑になる。たとえば食品添加物の発癌性のデータでは，白血球/リンパ球に限定すると有意に腫瘍を増加させるが，ある型の肝臓腫瘍の発生率を調べると，むしろ有意に減少するという（Haseman [06], Weinbergら [47]）。このようなとき，何を集約評価値にするか，有意水準をどう調整するかは大きな問題である（Haseman [07], Salsburg [34], Finkelsteinら [04], Bicksら [01]）。これは催奇形性において生じている問題と同種の問題で，まだこれについての明快な解答は存在しない。

7. 変異原性と遺伝毒性における問題

動物は遺伝子をととして種として生存している。遺伝を支配する遺伝子を根本的に変える性質は毒性の一つである。この毒性は，変異原性（mutagenicity）あるいは遺伝毒性（genotoxicity）として認識されている。

前臨床試験では，変異原性を短期試験で調べる。試験法は非常に多くあって，（1）染色体異常試験，（2）小核試験，（3）突然変異復帰試験，（4）不定期DNA合成試験，等到大別されている。

どの試験法でも，評価のための反応値（endpoint）は試験法ごとに定まっていて，データの型はたいいてい計数値である。実験はおおむね，陰性対照群，複数の用量群，陽性対照群の3～7群で行われ，定められた数のプレートの中のコロニー数，定められた数の観測細胞や赤血球の中での異常数，異常の頻度がデータとなる。短期試験であるから，実験の反復は比較的容易であり，実験の管理も長期試験に比べて容易である。このため十分注意すれば，陰性対照と陽性対照のデータで，実験ごとのばらつきを小さくできる。陰性対照には，モデルとして2項分布やポアソン分布が想定できることが多く，陽性対照には，負の2項分布が想定できることが多い。

変異原性試験のデータ解析については，英国の変異原学会がデータ解析の問題点を整理している（Kirkland [18]）。

一般に薬物の毒性は，用量を増すと反応値が大きくなることで存在が認識される。ところが変異原性の場合には，他の側面での毒性が，一つの試験法での変異原性の発現を妨害すること

がある。その典型がエームズ試験で、突然変異で生じるコロニーが、毒性の存在で抹殺され、用量とともに増加するはずのコロニー数が頭打ち(downturn)になることが知られている。変異原性試験のデータ解析では、この種の現象が存在することを常に考慮しなければならない(Simpsonら[37])。

統計学的側面では、①各群で、何匹の動物、何個のプレートを用意し、観察細胞等を何個にするのが標準的な試験法として適切か、②実験の用量としては、最大許容量(Maximum Tolerated Dose)をどのような視点で定めるのがよいか、③背景対照群(過去の実験での対照群のデータ)をどのように利用するのがよいか、④個々の試験法での変異原性陽性および陰性の判断としてどのような規準がよいか、⑤変異原性の総合判断に、どの試験法をどのように組み合わせ、どの種と性で試験を行うのがよいか、等が問題になる。Tennantら[42]はアメリカの国家毒性調査試験(National Toxicology Program: NTP)のデータに基づいて、予測が極めて難しいことを報告している。

8. 生殖性と催奇形性における問題

妊娠初期の睡眠薬が出生児に異常をもたらしたサリドマイド事件は、薬品にたいして、生殖に関する毒性の吟味が必要という認識をもたらした。この毒性は、注目する側面に応じて、発生毒性(developmental toxicity)、生殖毒性(reproductive toxicity)、催奇形性(teratologic toxicity)などと呼ばれるが、試験としては、生殖期にある雄または雌の母獣に薬物を投与し、胎児あるいは出生児の状態を観察するものである。医薬品毒性試験法ガイドライン(渡邊ら[07])では、妊娠前及び妊娠初期投与試験(Segment I)、胎仔の器官形成投与試験(Segment II)、周産期及び授乳期投与試験(Segment III)に分けている。

データとしては、妊娠黄体数、着床数、生存胎児数、形態異常が認められた胎児数、出生児の体重、などが観察される(白須ら[38])。その中で、①胎児の成長に関する偶然的変動には、親が同じかどうかによる影響すなわち同腹効果(litter effect)があるが、それをどのように評価するか(Kupperら[24], Krewskiら[21])、②多くの観測項目をまとめた特性量として何を用いるか、の二つが認識上及びデータ解析上で大きな問題となる。後者の問題では癌原性の場合と同様な対処法が模索されている。

9. 行動毒性における問題

毒性学の進歩によって、最近では、致死性あるいは臓器の明瞭な損傷ではない毒性が注目されるようになってきている。たとえば、記憶、注意力、学習能力、知覚機能、適応能力などへの影響である。一般にこの種の機能損傷を生じさせる性質を行動毒性(behavioral toxicity)という。

行動毒性の測定には、計量心理学でのいろいろな手段が利用できる。しかし前臨床試験では、対象が動物であるから、その手段がある程度限られると同時に結果の解釈に難しいところが少ない。たとえば餌と苦痛(たとえば電気刺激)を通しての学習行動が、薬物の摂取によって変化したかどうかは、統計的に把握できる。しかしその変化が、記憶についてのものか適応能力についてのものかは、生物学・心理学など他の分野の理論や知見でしか理解できない。統計学的には、変化の大きさの検出法だけしか問題にできない。

試験法を決めてしまえば、行動毒性のデータは一般にそれほど複雑でない。データは、対照群といくつかの用量の薬物投与群において行動を反復観測したものである。試験法は一般にそれほど単純でなく、測定を完全に自動化するのは難しいので、群の大きさは5匹あるいは10匹という程度の値である。

統計学的に問題になるのは、①反復観測 (repeated measures) の時間的非独立性を、どのように解析に取り入れるか、②用量と共に分散が大きくなりがちという不等分散性を、解析手法の選択にどのように考慮するか、③何を抛りどころにしてパラメトリック手法、ノンパラメトリック手法、変数変換の使い分けをするかである。

10. おわりに

前臨床試験には、実験の計画とデータ解析において、統計家の関与が求められる問題がいろいろ存在する。その多くは、新たに新規な理論や手法を用意することではなく、直面している問題を適切に定式化することで解決できるものと感じられる。しかし実際にそれを解決するには、実験の専門家と統計学者との間で、十分な知識の交換と共同作業が必要である。本稿では、統計家が統計学の雑誌に前臨床試験に関連してどんな統計手法を提案し、成果を上げているかという視点ではなく、現実の試験・実験のデータ解析でどのような統計学的問題があるかという視点で説明を行った。これは、その方が成果をあげる共同作業のきっかけになると考えたからである。

現在筆者のところには、①確率的に群分けしたにもかかわらず結果として個体の群間差が存在してしまったときにデータをどう解析すればよいか、②事前値を事後値の解析にどう利用すればよいか、③毒性が無視できるとして次のステップに進むために判断基準をどう設定すればよいか、④変異原性等の特定の毒性を調べるときに用量をどのような基準で設定するのが合理的か、⑤代替法の十分性はどのようにして判断すべきか、⑥創薬設計ではどのような方法論を用いるべきか、等多くの問題が寄せられている。このような問題について寄与を行う統計家により多く出現することを望みたい。

参 考 文 献

- [1] Bicks, M. & Krewski, D. (1989). Statistical issues in the analysis of the long-term carcinogenicity bioassay in small rodents, *Fundam. Appl. Toxicol.*, **12**, 202-221.
- [2] Crowder, M. J. and Hand, D. J. (1990). Analysis of repeated measures, Chapman and Hall.
- [3] Falk, R. W. (1989). Hommel's Bonferroni-type inequality for unequally spaced levels, *Biometrika*, **76**, 189-191.
- [4] Finkelstein, D. M. & Schoenfeld, D. A. (1989). Analysis of multiple tumor data from a rodent carcinogenicity experiment, *Biometrics*, **45**, 219-228.
- [5] Freedman, D. A. & Zeisel, H. (1988). The quantitative assessment of cancer risks from mouse to man, *Stat. Sci.*, **3**, 3-28.
- [6] Haseman, J. K. (1983). A reexamination of false-positive rates for carcinogenicity studies, *Fundam. Appl. Toxicol.*, **3**, 334-339.
- [7] Haseman, J. K. (1990). Use of statistical decision rules for evaluating laboratory animal carcinogenicity studies, *Fundam. Appl. Toxicol.*, **14**, 637-648.
- [8] Hayashi, M., Yoshimura, I., Sofuni, T. and Ishidate, M. (1989). A Procedure for Data Analysis of the Rodent Micronucleus Test Involving a Historical control, *Environmental Molecular Mutagenesis*, **13**, 347-356.
- [9] Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative, *Ann. Stat.*, **12**, 61-75.
- [10] Hochberg, Y. & Tamhane, A. C. (1987). Multiple comparison procedures, Wiley.
- [11] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, **75**, 800-802.
- [12] Hoekstra, J. A. (1991). Estimation of the LC50, A review, *Environmetrics*, **2**, 139-152.
- [13] Hoel, D. G. & Yanagawa, T. (1986). Incorporating historical controls in testing for a trend in proportions, *Jour. Amer. Stat. Assoc.*, **81**, 1095-1099.

- [14] Holland, B. S. & Copenhaver M. D. (1987). An improved sequentially rejective Bonferroni test procedure, *Biometrics*, **43**, 417-423.
- [15] Hommel, G. (1989). A comparison of two modified Bonferroni procedures, *Biometrika*, **76**, 624-625.
- [16] Hothorn, L. (1989). Robustness study on Williams and Shirley procedure with applications in toxicology, *Biometrical Jour.*, **31**, 891-903.
- [17] Hothorn, L. ed. (1991). Statistical methods in toxicology, Springer-Verlag.
- [18] Kirkland, D. J. ed. (1989). Statistical Evaluation of Mutagenicity Test Data, Cambridge Univ. Press.
- [19] 小島 肇, 花村朝夫, 佐藤 淳, 小西宏明, 吉村 功. (1992). 代替法試験の組合わせによる眼刺激性の予測, 日本動物実験代替法学会第6回大会予稿集, 126-127.
- [20] Krewski, D., Smithes, R. T. & Colin, D. (1987). Tests for trend in binomial proportions with historical controls: A proposed two stage procedure, *Biostatistics* (MacNeill, J. B. & Umphrey, G. J., eds.), Reidel Publishing, 61-69.
- [21] Krewski, D., Colin, D. & Yanagimoto, T. (1989). Statitlcal methods for the developmental toxicity studies, In Current Issues in Teratology (eds. Kehra, K. Grice, H. & Clegg, D.), Springer-Verlag, New York, 69-80.
- [22] Krewski, D. and Franklin, C. (1991). Statistics in Toxicology, Gordon and Breach Science Publishers.
- [23] 栗木 哲, 広津千尋(1990). 累積カイ二乗の最大成分に基づく多重比較法一有意確率計算と用量水準比較への応用一, 応用統計学, **18**, 129-141.
- [24] Kupper, L. L., Portier, C., Hogan, M. D. & Yamamoto, E. (1986). The impact of litter effects on dose-response modeling in teratology, *Biometrics*, **42**, 85-98.
- [25] Lee, E. T. (1992). Statistical mehtods for survival data analysis, 2nd edition, Wiley.
- [26] Lehmaccher, W., Wassmer, G. & Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate, *Biometrics*, **47**, 511-521.
- [27] Margolin, M. H. & Risko, K. J. (1984). The use of historical data in laboratory studies, Proc. Int. Bio. Conf., 21-30.
- [28] 松田眞一, 永田 靖(1990). 多重比較法における新たな検出力の提案と各手法の比較, 応用統計学, **19**, 93-113.
- [29] 森田智視 (1991). 比較毒性試験における多変量解析の応用, 東京大学医学部保健学科卒業論文.
- [30] Moore, D. F. (1987). Modelling the extraneous variance in the presence of extra-binomial variation, *Appl. Stat.*, **36**, 8-14.
- [31] Mudholkar, G. & McDeromott, M. P. (1989). A class of tests for equality of ordered means, *Biometrika*, **76**, 161-168.
- [32] Peace, K. E. ed. (1988). Biopharmaceutical Statistics for Drug Development, Marcel Dekker. (邦訳, 中上節夫, 森川敏彦 監訳 (1992). 医薬統計学, サイエンティスト社.
- [33] Peto, R., Pike, M. C., Day, N. E., Gray, R. G., Lee, P. N., Parish, S., Peto, J., Richards, S. & Wahrendorf, J. (1980). Guidlines for simple sensitive significance test for carcinogenic effects in long-term animal experiments, IARC Monographs, Suppl. 2 (Lyon, France).
- [34] Salsburg, D. (1989). Does everything "cause" cancer: An alternative interpretation of the "carcinogenesis" bioassay., *Fundam. Appl. Toxicol.*, **13**, 351-358.
- [35] Sawyer, C., Peto, r., Bernstein, L. & Pike, C. M. (1984). Calculation of carcinogenic potency from long-term animal carcinogenesis experiments, *Biometrics*, **40**, 27-40.
- [36] Selwyn, M. R. (1989). Dual controls, p-value plots, and the multiple testing issue in carcinogenicity studies, *Envir. Health Perspectives*, **82**, 337-344.
- [37] Simpson, A. & Margolin, R. H. (1986). Recursive nonparametric testing for dose-response relationship subject to downturns at high doses, *Biometrika*, **73**, 589-596.
- [38] 白須泰彦, 吐山豊秋編 (1985). 新毒性試験法一方法と評価一, エル・アイ・シー.
- [39] Tamura, N. & Young, S. S. (1986). The incorporation of historical control information in tests of proportions: Simulation study of Tarone's procedure, *Biometrics*, **42**, 343-349.
- [40] Tarone, R. E. (1982). The use of historical control information in testing for a trend in proportions, *Biometrics* **38**, 215-220.
- [41] Tattersall, M. L. (1982). Statistics and the LD50 study, *Arch. Toxicology*, suppl. 5, 267-270.

- [42] Tennant, R. W., Margolin, R. H., Shelby, M. D., Zeiger, E., Haseman, J. K., Spalding, J., Caspary, W., Resnick, M., Stasiewicz, S., Anderson, B. & Minor, R. (1987). Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assays. *Science*, **236**, 933-941.
- [43] 統計数理研究所共同研究リポート 18 (1989). 多重比較方式の諸問題.
- [44] 統計数理研究所共同研究リポート 22 (1990). 毒性・薬効データ解析における多重比較法の利用.
- [45] Tukey, J. W. (1991). The philosophy of multiple comparisons, *Stat. Science*, **6**, 100-116.
- [46] 渡邊 徹, 堀内茂友編 (1991). 改正医薬品毒性試験法ガイドライン・GLP 基準, 地人書館.
- [47] Weinberg, A. M. & Storer, J. B. (1985). Ambiguous carcinogens and their regulation, *Risk Analysis*, **5**, 151-156.
- [48] Williams, D. J. (1986). Interval estimation of the median lethal dose, *Biometrics*, **42**, 641-646.
- [49] Wright, S. P. (1992). Adjusted p-values for simultaneous inference, *Biometrics*, **48**, 1005-1013.
- [50] Yanagawa, T., Hoel, D. G. & Books, G. T. (1989). A conservative use of historical data for a trend test in proportions, *J. Jpn. Stat. Soc.*, **19**, 83-94.
- [51] Yoshimura, I. & Matsumoto, K. (1993). Notes on the Use of Historical Controls, *Envir. Health Perspectives*, (to appear).
- [52] 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.
- [53] 吉村 功, 大橋靖雄 編 (1992). 毒性試験データの統計解析, 地人書館.
- [54] 吉田道弘 (1991). 多重比較, 日本科学技術連盟. (本文献は「医薬データの統計解析専門コース」のテキストで一般の入手は困難であるが, その内容の多くは吉田道弘・永田靖共著で 1994 年に発行予定の書物に含まれるものと思われる).
- [55] Zbinden, G. & Flury-Roversi, M. (1981). Significance of the LD50-test for the toxicological evaluation of chemical substances, *Arch. Toxicology*, **47**, 77-99.