一般毒性試験における外れ値と用量反応パターンの統計学的な評価に関する研究

浜田知久馬

①

一般毒性試験における外れ値と用量反応パターンの統計学的な評価に関する研究

浜田　知久馬

目次

1．研究目的・研究の背景

本論文では，マウス・ラットなどのげっ歯類を用いた反復投与毒性試験における，一元配置型データの統計解析の問題を取り上げ，特に標準的な方法が確立されてない，外れ値と用量反応パターンの統計学的な評価について，研究した結果を述べる．

1．1　毒性試験の統計解析の特徴

新薬の開発にあたっては臨床試験に入る前に，薬剤の有害性を確認するため，マウス・ラットなどのげっ歯類，イヌ・サルなどの大動物を用いた毒性試験が義務づけられている．毒性試験は薬理試験や臨床試験とは異なり，生体での異常反応を検出することが目的であり，通常は薬効量の数十倍から数百倍の投与量が設定される．したがって，個体差も強く現れ，必然的にバラツキも大きくなるが，動物愛護の観点からも使用匹数は限られており，むやみに増やすこともできない状況であり，限られた資源の中で，精度の高い評価が要求されている．毒性試験によって生じるデータの統計解析を行う上で，留意しなければならない点についてまとめてみた．

1）実験計画の類似性

試験の実験計画については，評価項目，群数，サンプルサイズ等がガイドラインで実質的にはほぼ規定されており（厚生省薬務局審査第一課(1991)），試験計画のバリエーションは臨床試験と比べるとかなり小さい．群数は対照群を含めて4～5群，サンプルサイズは，マウス，ラットなどを用いたげっ歯類の試験では，1群あたり10～20匹，イヌやサルを用いた大動物の試験では，3～5匹が慣習的に用いられている．通常，群分けの際には，体重によってブロック化する層別無作為化割り付けがなされる．

2）評価項目(エンドポイント)が多数存在する

一般毒性試験では試験物質の有害性を確認することが目的の試験であるため，本質的にエンドポイントが多数存在する．それぞれの個体について体重，摂餌量，飲水量，血液学的検査，血液生化学的検査，尿検査，臓器重量（絶対・相対値）などが測定され，通常1つの実験の中で数百項目以上の多数に渡って検定がなされる．これらのデータから試験物質の投与と関連を持ち，人に対する有害性が予想される変化を抽出することが必要となる．有害性のスクリーニングが目的であるため，プライマリーエンドポイント等の特別な指標は通常設定されない．

3）評価項目の分布形が多様である．

評価項目の数が多いのに加え，その分布形も多様である．体重あるいは臓器重量は，多くの場合，近似的に正規分布とみなすことができる．これに対しALT, ASTに代表されるような生化学検査値は，右に歪んだ分布であり，対数変換した方が正規分布にしたがう場合が多い．また，臨床所見あるいは生死の有無などは2値データであり，病理所見などは順序カテゴリカルデータになる．

4）非等分散あるいは外れ値が存在する場合がある．

薬物の用量の増加に伴い，反応の平均値のみならず，バラツキも増大するのが典型的な

毒性反応としてしばしば観測される．このため多くの統計手法が前提とする等分散性が成立しない場合がある．また他の観測値と比べて著しく隔たった値（外れ値:outlier）が頻繁にみられる．外れ値が生じる理由としては，測定ミス，個体の異常，個体に特異的な毒性等などが考えられるが，多くの場合原因を特定するのは困難である．また外れ値が存在すると，結果の評価が困難になることがある．このように統計解析手法を選択するにあたっては，非等分散性，外れ値に適切に対処する必要がある．

5）曲線的な用量反応関係の可能性

通常は，用量の増加に伴い，反応が単調に増加または減少することが期待される．しかしながら，まれに用量反応曲線が高用量で頭打ちしてしまうことがある．このような現象はdownturnと呼ばれ，Ames試験や小核試験などの変異原性試験では，頻繁に観測される現象である．

6）統計の非専門家による評価

欧米では各製薬企業の非臨床部門あるいは受託試験所に，統計の専門家が配置されており，試験計画・解析・レポーティングの段階でサポートを行っているが，我が国では，生物統計家の絶対数が不足しており，非臨床分野の統計解析の専門家は産・官・学を通じて皆無といってもよく，実際の評価は，統計の専門教育を受けていない実験者の手によって行われている．したがって数学的に極端に難しい手法や，複雑な手法の利用は避けるべきである．

1．2 ツリー型アルゴリズム

我が国では，少なくとも数年前までは，多くの製薬企業あるいは受託研究機関において，Figure 1に示したようなツリー型アルゴリズム(山崎等(1981))を用いて，毒性試験データの評価が行われてきた．

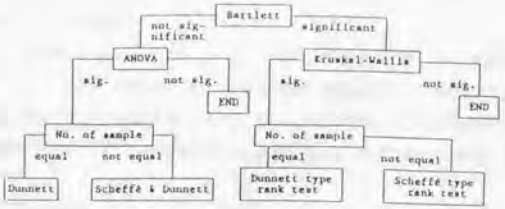


Figure 1　Standard tree algorithm in toxicity studies

このアルゴリズムでは最初に予備検定として，データの等分散性についてBartlett検定を用いて検討し，その結果に応じてパラメトリックとノンパラメトリック手法の使い分けを決定する．また対照群と各用量群の比較について，例数が群間で等しいか，等しくないかによって，それぞれDunnett法とScheffe法を使い分けている．毒性試験では評価項目数が多いため，データの分布形を，実験前に決定することが困難であり，パラメトリック手法とノンパラメトリック手法の選択を，得られたデータに依存して決定することは，実用的であった．またそれまで，検定の多重性の問題を無視して，t検定を繰り返し適用してきた毒性試験の統計解析に，多重比較の手法を導入した歴史的な意義は大きいといえるが，同時に様々な批判を受け続けてきた（Kobayashi, Watanabe and Inoue(1995)，浜田・岸本(1995)）．このツリー型アルゴリズムの問題点は，次のようにまとめることができる．

1） 例数がアンバランスな場合のScheffe法の使用

このアルゴリズムの最も重大な問題点は，例数がアンバランスな場合にScheffeの方法を用いる点である．Dunnett法は原典では例数が等しいことを前提として提案された手法であるが，最近ではアンバランスなケースについても精度の高い近似式が知られており，また統計パッケージSASのPROBMC関数(SAS Institute(1996))を用いれば，例数が群間で異なっていても，正確にp値を計算することが可能である．対照群と各用量群の対比較の問題にScheffe法を適用すると，過度に保守的になり，極端に検出力が低下する（吉村・大橋(1992)）．特に毒性試験でアンバランスになるケースは，高用量で毒性による死亡が起きている場合が多く，このように有害反応の発現が予想されるケースで，必要以上に保守的な方法を用いるのは明らかに好ましくない．

2）分散分析と多重比較の併用

このアルゴリズムでは分散分析で有意な結果が得られた場合のみ，多重比較を行うことになっているが，本来この2つの手法は独立に行うべきものであり，このように両手法が有意になったときのみ有意とすると，実際の有意水準は，それぞれの検定の有意水準より小さく，保守的になる．

3）Bartlett検定による予備検定

このアルゴリズムでは，Bartlett検定によって等分散性を検定し，その結果に基づいて，パラメトリック・ノンパラメトリック手法を使い分けているが，2種類の手法の使い分けは本来分布形に基づいてなされるべきである．両手法とも等分散性の前提条件（比較する2群の間で分布形が等しい）は必要であり，Bartlett検定を予備検定に用いるのは論理的に矛盾している．

4）外れ値についての検討

外れ値はデータ解析の結果にしばしば重大な影響を与えるので，初期の段階で検出しておく必要がある．また特に毒性試験では，異常に飛び離れた個体の存在自体が，毒性発現の重要なインフォーメーションになる場合がある．したがってデータ解析の際は，外れ値について検討する必要があるが，このツリー型のアルゴリズムでは，外れ値についての検

討が行われていない．

5）用量相関性・用量反応パターンについての検討

毒性試験の重要な目的は，被験物質が有害性を有するかどうかを確認することである．このため対照群と比べて変化のある用量群を見つけるとともに，用量反応のパターンを評価し，被験物質に関連した毒性の知見や生体内動態，代謝などの知見を統合して有害性が予想される変化を見つけ出す必要がある．このアルゴリズムでは対照群との比較として，Dunnett 法あるいは Scheffe 法が適用されるが，用量反応のパターンについての検討がなされていない．

1．3 最近の毒性試験の統計解析の動向

医薬品の開発過程を日米欧の3極でハーモナイズする動き ICH(International Conference on Harmonization)と関連して，毒性試験の統計解析について，最近急速な変化がみられる．非臨床試験の統計解析の問題について，1996年3月にベルギーにおいて，第1回の DIA(Drug Information Association)のワークショップ Statistical Methodology in Non-Clinical and Toxicological Studies が開催された．このワークショップの目的は，製薬企業，教育研究機関，行政から統計学の専門家を招き，毒性試験あるいは ICH に関連したドキュメントで記載するのに適切な統計手法に関して，議論することであった．また，同年8月に東京で開催された第3回の DIA の東京ワークショップ(Annual Biostatistics Meeting in Tokyo)にて，非臨床試験の統計解析のガイドラインについてのセッションが設けられた．またこの東京ミーティングでの成果を受け，毒性試験の統計について，教育研究機関，行政，製薬企業が共同してリコメンデーションがなされた(Hothorn et al. (1997))．このリコメンデーションは毒性試験の目的，試験計画，実施，解析，解釈について統計学的に考慮すべき点をまとめたものである．この論文では毒性試験の統計解析について，どのような考え方で行うべきかが示されている．ただし個々の場面で具体的にどの手法を用いるかが，クックブック的に示されているわけではない．適切な統計手法は時代の流れに応じて変化するので，個々の手法の選択は統計の専門家の判断にまかせられている．このため標準的な手法については，各国で独自に確立する必要がある．

このような近年の国際化への動きに伴い，我が国においても非臨床試験の統計学的な評価の重要性についての認識は高まりつつある．臨床試験については，人種差，医療システム等の違いの問題があるため，異なった国で行われた試験結果をどこまで共有化できるかについては議論のあるところであるが，これと比較して，非臨床試験は，各国で同じ動物を用いて，同一の手技で試験を行うため，ある意味ではハーモナイズがしやすい領域であるといえる．またハーモナイズすることは，開発期間・コスト・実験に用いる動物数の節約などの様々なメリットをもたらす．欧米の製薬企業には非臨床の部門にも統計の専門家が配属されているのに対し，我が国では，臨床には統計の部門ができはじめているが，非

臨床部門については，統計の専門家は皆無に近い状態である．このため，急速な変化に毒性試験の現場が対応しきれず，混乱を起こしているのも事実である．例えば新薬を開発するにあたり要求される毒性試験の種類が多いため，ある化合物についての一連の毒性試験の全てを単独の製薬企業で行うことは困難であり，このうちのいくつかは，受託研究機関に外注するのが一般的となっているが，受託機関はその施設独自の統計解析プログラムを作成しており，新薬の申請に当たって一連の試験結果をまとめてみると，試験ごとに統計手法がまちまちで，統一性がとれないなどの問題が生じている．また最近では対照群との比較だけではなく，安全性評価の上でリスク評価に必要な用量相関性の解析が，新薬の申請にあたって要求されることが多くなっているが，適切ではない解析例があることが指摘されている(吉村(1995))．

1．4　研究目的

本研究ではこのような背景をふまえて，毒性学の専門家と共同して一般毒性試験において，標準的な統計解析手法を推奨することを目的とした．計量データを対象として，特に標準的な統計手法が確立されてない用量反応パターンの解析と外れ値の検出の問題を中心に検討した．統計解析の結果は，毒性家に直観的に受け入れられるべきであると考え，よく用いられる標準的な統計解析手法と毒性家の判断の一致度を評価し，毒性家のデータ評価のストラテジーの一端を明らかにした．その結果に基づき，毒性試験データの望ましい統計解析手法について考察した．また特に用量反応パターンの評価については，最大対比法の利用を提案し，その性能について評価した．

II．方法およびデータ

1）対象データ

ラット雌雄を用いた18の一般毒性試験データ（血液・生化学・臓器重量（絶対値・相対値））を入手し，本研究用のデータベースを作成した．延べ項目数は2001，用いたラット数は合計1711匹（雌:862匹，雄:849匹）になった．各試験で対照群を含めた群数は3-6群であり，1群あたりのサンプルサイズは10～25匹であった．

2）評価者

10年以上の毒性試験の評価の経験を持つ毒性学の専門家8人をノミネートした．8人中6人は毒性学の専門家であり，残りの2人は病理学者であった．また3人は獣医師の免許を有していた．それぞれの試験について4人づつ無作為に割り当て，8人の評価者が各項目の測定値のグラフ・生データ・平均値に基づいて，外れ値，用量相関性，変化を起こしている用量の判断を独立に行った（付録1参照）．研究番号と評価を行った毒性家の対応を次に示した．

研究番号	評価者番号				研究番号	評価者番号			
1(4)	8	6	3	7	2(5)	8	1	2	
3(4)	5	2	1	7	4(5)	2	6	7	5
5(5)	7	4	3	5	6(4)	8	4	7	5
7(3)	8	3	2	7	8(5)	7	5	6	2
9(5)	6	5	1	4	10(4)	6	7	2	5
11(6)	2	1	8	4	12(4)	1	7	5	
13(5)	5	7	1		14(4)	7	5	3	
15(3)	4	8	5	2	16(4)	1	8	7	3
17(5)	1	7	8	4	18(3)	7	5	8	

＊ ()内は群数を表す

計画ではすべての試験を４人づつ評価する予定であったが，期限までに評価が終了してないケースがあり，３人で評価した研究が５つ存在する．延べ評価項目数は，7485項目・人となった．

3) 統計学的な評価

いくつかの標準的な統計手法を，これらのデータに適用した．計算にあたっては統計解析パッケージ SAS を用いた．()内に用いた SAS のプロシジャ名を示した(SAS Institute(1996))．

1 外れ値の評価：skewness(UNIVARIATE)，kurtosis(UNIVARIATE)，studentized residual(GLM)を適用した

2 用量相関性の検定：回帰分析(REG)，対数変換後の回帰分析(REG)，ヨンキー検定(CORR)を適用した．

3 対照群と各用量群の比較：ノンパラ型の Dunnett 検定(PROBMC 関数を利用したマクロを作成した) を適用した．

なお外れ値については各群ごとにその有無を評価した．

・ skewness と kurtosis

skewness(歪み)と kurtosis(尖り)(吉村・大橋(1992))はモーメント統計量の概念によって導かれ，データが正規分布にしたがうとき，これらの統計量の期待値は0になる(kurtosisは定義によっては期待値が３となるが，ここでは０になる方の定義を用いた)．外れ値が存在すると，skewnessは絶対値が，kurtosisは値が正の方向で大きくなる．したがってこれらの統計量は，しばしば外れ値を検出するために利用される．それぞれの統計量が次の臨界値を越えたとき，有意水準αで当該群に外れ値が存在すると判断した．

$$|\text{skewness}| > Z_{\alpha/2} \cdot \text{SQRT}(6/N) \quad (1)$$

$$\text{kurtosis} > Z_{\alpha} \cdot \text{SQRT}(24/N) \quad (2)$$

ここでNはそれぞれの群当たりのサンプルサイズで，Z_{α}は正規分布の上側α%点である．

・studentized residual

studentized residual(竹内(1989))は，モデルが複数の説明変数を含む場合について，あてはまりのよくない観測値（外れ値）を検出するために導かれた統計量であるが，より単純なケースである一元配置分散分析では，次のように定義される．

$$Sr_{ij}=(Y_{ij}-Y_i^*)/SD^*_{pool} \qquad (3)$$

Sr_{ij}：i番めの群のj番めの観測値のstudentized residual

Y_{ij}：i番めの群のj番めの観測値の値

Y_i^*：Y_{ij}を除いた第i群の平均値

SD^*_{pool}：Y_{ij}を除いて，群間でプールした群内分散の平方根をとったもの

studentized residualの計算にあたっては，SASのGLMプロシジャを用いた．誤差分布が独立で等分散の正規分布にしたがうとき，それぞれのstudentized residualは独立に誤差平方和に相当する自由度を持つt分布にしたがうことが知られている．したがって，次の条件が満たされた場合，群iに有意水準αで外れ値が存在すると判定した．

$$Max(|SR_{ij}|) > t_{\alpha/(2\times N)}$$

$Max(|SR_{ij}|)$：群iにおけるstudentized residualの絶対値の最大値

$t_{\alpha/(2\times N)}$：誤差平方和に相当する自由度を持つt分布の上側α／(2×N)%点

両側有意水準α/2をNで割るのは，観測値の多重性についてBonferroniの不等式を用いた調整である．

・回帰分析，対数変換後の回帰分析，ヨンキー検定

用量相関性の検定としていくつかの手法が知られているが，ここでは代表的な3種類の方法を適用した．回帰分析は誤差分布に正規分布を前提として回帰直線を求め，傾きが有意に0と異なるか検定するパラメトリックなアプローチである．この方法と比較するため，対数変換した後の回帰分析・ノンパラメトリックなヨンキー検定を適用した．またこれらの用量相関性の検定との関連を調べるため，対照群と各用量群の比較について，ノンパラ型のDunnett検定を適用した．この検定を実施するにあたっては，SASのPROBMC関数を利用してjoint-ranking型のプログラムを作成した(浜田・岸本(1995))．このプログラムを用いれば，群間でサンプルサイズのアンバランスがあっても正確なDunnett検定のp値を計算することができる．ノンパラメトリックな多重比較法としては，Steel検定もseparate-ranking型の方法として知られているが，我が国の毒性試験の現場ではポピュラーではないので，採用しなかった．なおp値はいずれも両側で計算した．

Ⅲ．毒性家の判断と統計解析の一致度

3．1 外れ値の検出

・毒性家の判断と統計手法の一致度

skewness, kurtosis と studentized residual の3種の統計量を，外れ値を検出するために適用した．結果を得られたp値によって，p>0.05，0.05>p>0.01，0.01>p>0.001，0.001>p>0.0001，0.0001>p>0.00001，0.00001>pの6通りに分類し，毒性家の判断と比較した．Table 1-3は統計的な手法と毒性家の判断の関連を示している．毒性家ごとの外れ値の陽性判定率および，8人の毒性家をまとめた結果("all")を示している．

Table 1は skewness の結果を示している．p値>0.05 のときは，全体での陽性判定率(0.9%)は1%より小さく，また8人の毒性家間では0.1-2.6%の範囲にあった．統計学的な有意性が強くなるに連れ，陽性判定率は，上下しながら増加し，単調には変化してない．またp<0.00001と最も強い有意性がある場合でも，陽性判定率は36.5%に過ぎない．このように統計学的な結果と毒性家の判断の一致度が低かった理由については，後で例解する．更に評価者によって，外れ値の判定の仕方がかなり異なることがわかる．例えば毒性家1は，最も保守的で陽性と判定しにくいのに対し，毒性家6と8は革新的で陽性と判定しやすい．

kurtosis と毒性家の判断の一致度をTable 2に示した．5%水準で有意差がない場合は，全体での陽性判定率は0.8%で，毒性家間では0.1-2.5%の範囲でばらついた．この点についてはskewnessとほぼ同様な結果が得られているといえる．skewnessと比べると，有意な場合については，結果の改善がみられ，少なくとも全体での結果については，有意性が強くなるにつれ，陽性判定率は単調に増加している．しかしながらp<0.00001と最も強い有意性がある場合でも，陽性判定率は40%以下である．

studentized residual と毒性家の判断の一致度をTable 3に示した．5%水準で有意差がない場合は，全体での陽性判定率は0.3%で，毒性家の間での範囲は0.1-0.9%となった．

Table 1　Evaluation for outliers using skewness

Judgment of toxicologis	Result of statistical test Positive percent (number of evaluated item･groups)					
	p>0.05	0.05>p>0.01	0.01>p>0.001	0.001>p>0.0001	0.0001>p>0.00001	0.00001>p
1	0.1 (3577)	2.1 (235)	15.8 (114)	28.1 (32)	16.7 (24)	19.5 (41)
2	1.0 (3668)	6.4 (234)	18.6 (118)	18.5 (27)	30.0 (30)	54.4 (46)
3	0.5 (2058)	6.0 (83)	32.1 (28)	70.0 (10)	20.0 (5)	(0)
4	2.3 (2881)	17.3 (139)	44.3 (61)	77.8 (9)	9.5 (21)	0.0 (7)
5	0.3 (5202)	5.2 (306)	14.8 (142)	55.6 (18)	33.3 (21)	21.1 (19)
6	2.6 (2865)	21.0 (162)	34.5 (84)	50.0 (14)	58.3 (12)	21.1 (19)
7	0.1 (6019)	4.0 (325)	22.5 (138)	50.0 (20)	29.2 (24)	33.3 (12)
8	1.7 (3710)	16.4 (201)	42.7 (96)	48.2 (27)	14.8 (27)	58.8 (34)
all	0.9(29980)	8.6(1685)	25.4 (781)	43.3(157)	25.0(164)	36.5(178)

Table 2 Evaluation for outliers using kurtosis

Judgment of toxicologis	Result of statistical test Positive percent(number of evaluated item･groups)					
	p>0.05	0.05>p> 0.01	0.01>p> 0.001	0.001>p>0 .0001	0.0001>p> 0.00001	0.00001>p
1	0.1 (3547)	1.7 (173)	0.9(107)	9.8 (61)	24.1 (29)	25.5(106)
2	0.9 (3647)	5.3 (169)	8.4(107)	22.2 (54)	21.6 (37)	39.5(109)
3	0.2 (2004)	7.3 (82)	9.8 (51)	31.6 (19)	36.4 (11)	47.1 (17)
4	1.9 (2831)	17.4 (121)	20.3 (69)	32.4 (34)	41.2 (17)	37.0 (46)
5	0.2 (5106)	4.0 (252)	5.8(155)	13.2 (68)	14.9 (47)	32.5 (80)
6	2.5 (2833)	10.4 (135)	28.4 (74)	34.4 (32)	46.4 (28)	44.4 (54)
7	0.1 (5894)	1.1 (272)	3.6(166)	16.1 (81)	22.9 (48)	41.6 (77)
8	1.4 (3666)	12.3 (154)	25.0 (96)	39.0 (59)	46.2 (26)	48.9 (94)
all	0.8(29528)	6.3(1358)	10.8(825)	22.3(408)	28.4(243)	38.3(583)

Table 3 Evaluation for outliers using studentized residual

Judgment of toxicologis	Result of statistical test Positive percent (number of evaluated item･groups)					
	p>0.05	0.05>p> 0.01	0.01>p> 0.001	0.001>p>0 .0001	0.0001>p> 0.00001	0.00001>p
1	0.1 (3466)	0.0 (240)	2.7 (150)	10.2 (59)	10.7 28)	41.3 (80)
2	0.7 (3545)	3.3 (243)	8.4 (143)	11.4 (70)	32.0 25)	53.6 (97)
3	0.0 (1919)	2.9 (137)	8.3 (72)	20.0 (25)	25.0 (8)	69.6 (23)
4	0.6 (2738)	14.0 (172)	24.8 (113)	41.7 (36)	70.6 (17)	69.1 (42)
5	0.0 (4909)	0.8 (366)	3.3 (215)	14.1 (85)	22.9 (35)	42.9 (98)
6	0.9 (2749)	13.2 (197)	28.9 (97)	52.1 (48)	52.9 (17)	85.4 (48)
7	0.0 (5661)	0.7 (415)	0.9 (232)	5.6 (89)	8.6 (35)	52.8(106)
8	0.4 (3541)	9.7 (228)	24.1 (141)	45.2 (62)	34.8 (23)	69.0(100)
all	0.3(28528)	4.5(1998)	10.4(1163)	21.9(474)	28.2 (188)	56.9(594)

これはskewnessあるいはkurtosisの結果に比べると低く，5%水準で有意でない場合に陽性と判断されることは非常に少ないといえる．有意であるケースについては，統計的な有意性が強まるに連れ，単調に陽性判定率は上昇する．特に $p<0.00001$ と最も強い有意性がある場合は，陽性判定率は56.9%で，これはskewness，kurtosisの結果に比べるとかなり高い値である．

･統計手法間で結果が食い違うケース

Table 4 Raw data of total bilirubin (mg/dl)

Group	Dose	Raw data(mg/dl)									
Control	0mg	0.11	0.10	0.03	0.10	0.10	0.09	0.10	0.08		
low-dose	1mg	0.08	0.12	0.11	0.11	0.10	0.12	0.10	0.11	0.11	0.10
mid-dose	3mg	0.13	0.11	0.13	0.16	0.14	0.17	0.16	0.17	0.16	0.12
high-dose	10mg	0.16	0.28	0.27	0.29	0.33	0.33	0.30	0.30	0.27	

Table 5　Summary statistics of total bilirubin (mg/dl)

Group	Dose	Mean	SD	Skewness	Kurtosis	Max	SMAX*	Min	SMIN**
Control	0mg	0.089	0.025	−2.208	5.267	0.11	0.7	0.03	−2.2
low-dose	1mg	0.106	0.012	−1.072	1.854	0.12	0.5	0.08	−0.9
mid-dose	3mg	0.145	0.022	−0.325	−1.475	0.17	0.9	0.11	−1.2
high-dose	10mg	0.281	0.051	−1.899	4.693	0.33	1.8	0.16	−6.1

* studentized residual of maximum value　** studentized residual of minimum value

これらのモーメント統計量に基づく外れ値の判定と，毒性家の判断の一致度が高くない理由をTable 4のデータを用いて例解する．このデータ（ビリルビン）は，研究用データベースの中から抽出したものである．散布図をFigure 2に，生データ，要約統計量と外れ値を検出するための統計量を，それぞれTable 4,5に示した．skewnessとkurtosisの値から，対照群と最高用量群に外れ値の存在が示唆される．この2つの群では，他の観測値と比べて，相対的に低めに外れた値0.03（対照群），0.16（高用量群）が存在するため，skewnessとkurtosisが大きな値をとる．しかしながら高用量群のSDは対照群のほぼ2倍であり，Figure 2が示しているように，高用量群の外れ具合の方が対照群のそれよりかなり大きい．したがってskewnessとkurtosisの値自体は，対照群の方が少し大きいにもかかわらず，毒性家は高用量群にのみ外れ値があると判断する．毒性家は外れ値を検出するため，データの分布形のみならず，バラツキの大きさも考慮しているといえる．

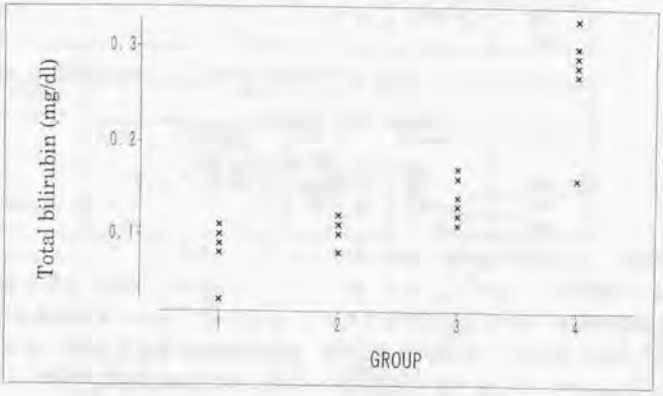


Figure 2　Scatter plot of total bilirubin (mg/dl)

ほとんどの観測値が非常に狭い範囲に偶然的に集中し，ある観測値が少し離れたところに位置すると，例えその離れ方の程度がとるに足らなくても，skewnessとkurtosisの値は大きくなる．これが毒性家の判断と，これらの統計量を利用した評価の食い違う主な原因である．一方studentized residualは，高用量群の0.16に対しては6.1で，これは対照

群の0.03に対する-2.2と比べるとかなり大きい．studentized residualは，平均値からの距離を，群間でプールした分散の平方根で割ることによって基準化される．このように共通した分散を用いることにより，studentized residualではより安定した結果が得られる．

3．2　用量相関性の検定

・毒性家の判断と統計手法の一致度

Table 6は，ヨンキー検定の毒性家の判断の一致度を示した．統計学的な有意性は，p>0.05，0.05>p>0.01，0.01>p>0.001，0.001>pの4段階に分類した．評価者×項目数の累計は7485になった．p値が0.05以上で有意でない場合（4962項目・人）について，毒性家が陽性と判断した割合は3.85%であった．統計学的な有意性が強くなるにつれ，陽性判定率も徐々に増加するが，統計学的には最も強い有意性がある0.001>pのケースでも，毒性家の陽性判定率は，64%に過ぎなかった．この理由については後で説明する．

Table 6　Consistency of dose dependency.

Decision of toxicologists	Result of Jonckheere test Frequency (Column percent)				
	p>0.05	0.05>p>0.01	0.01>p>0.001	0.001>p	Total
negative	4771 (96.15)	611 (83.02)	363 (70.49)	454 (35.69)	6199
positive	191 (3.85)	125 (16.98)	152 (29.51)	818 (64.31)	1286
Total	4962	736	515	1272	7485

Table 7　Consistency of change at maximum dose

Decision of toxicologists	Result of Dunnett type test Frequency (Column percent)				
	p>0.05	0.05>p>0.01	0.01>p>0.001	0.001>p	Total
negative	5112 (90.93)	294 (50.17)	129 (26.38)	85 (10.79)	5620
positive	510 (9.07)	292 (49.83)	360 (73.62)	703 (89.21)	1865
Total	5622	586	489	788	7485

Figure 3は，ヨンキー検定に回帰分析と対数変換後の回帰分析を加えて，3種類の統計手法と毒性家の判断の一致度を示している．図をみると，いずれの用量相関性の検定でも一致度はほぼ同様であることがわかる．p値が0.05～0.01のときは，陽性判定率は20%以下であり，毒性家の用量相関性の判断は，統計学的な5%水準より保守的である．平均すると，3手法間の成績に大差はなかったが，中には大きく食い違うケースもあった．これについては次節で検討するが，外れ値が存在すると結果が大きく食い違う場合があった．この節では以下，外れ値に頑健であったヨンキー検定の結果を示す．

それぞれの試験で用いられた最高用量での変化について，Dunnett型検定と毒性家の判断との一致度をTable 7に示した．用量相関性の検定と比較すると，毒性家の判断との高い一致度が示された．特にp値が0.001と最も強い有意差がある場合には，陽性判定率は約90%となった．

Figure 4にそれぞれの用量ごとの陽性判定率を示した．p値が0.05～0.01のときは，用量が高くなるに伴い，陽性判定率が高くなる傾向があった．この理由は，毒性家が，低い用量でp値がぎりぎり，5%水準で有意となる微妙な変化を見つけた場合，より高い用量で変化が起きているかを調べ，起きていないようであれば，低い用量でも変化なしと判断する傾向があったためである．しかしながらp値が0.01より小さくなる強い変化については，用量群間で大きな差は見られなかった．

Figure 5は，それぞれの毒性家ごとに用量相関性の判定結果を示したものである．図中で"All"は8人の毒性家をまとめた結果を表している．毒性家間で判定に大きな違いがあることがわかる．例えば6番の評価者は最も革新的であり，陽性と判断しやすい傾向がある．p値が0.05より大きく有意でないケースであっても，15%の確率で陽性と判定している．

対照的に，1番と8番の毒性家は用量相関性の判定が保守的であり，p値が0.05より大きいときには全く陽性と判定せず，p値が0.001以下と統計学的には明らかな有意差がある場合でも，陽性判定率はわずかに35%である．残りの5人の評価者はこれらの中間型である．このように統計学的検定と毒性家の判断の一致度が低かったのは，毒性家間の判定の大きなバラツキのためである．

一方，最高用量における変化について，毒性家ごとの判定結果をFigure 6に示した．p値が0.05～0.01と，統計学的にぎりぎりで有意になる場合には，毒性家間で陽性判定率がばらつくことがわかる．しかしp値が0.001以下になると変動の大きさは小さくなる．最も保守的な毒性家8でも，このケースでは80%の確率で陽性と判定している．用量相関性の結果と比較すると，評価者間のバラツキは小さいといえる．評価者によって若干判断基準が異なるものの，その違いは質的なものではない．

Table 8 Relation of judgments between D-R and change at maximum dose

Person	1-MaxD negative Dose-Response		1-Maxd positive Dose-Response		
Freq.	negative	positive	negative	positive	Total
1	643	0	125	65	833
2	768	2	15	134	919
3	409	0	2	135	546
4	392	2	98	168	660
5	956	1	81	284	1322
6	459	1	10	214	684
7	1188	0	143	205	1536
8	798	1	112	74	985
total	5613	7	586	1279	7485

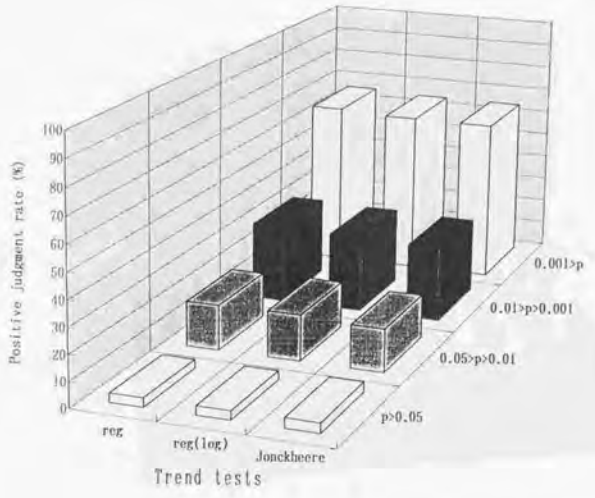


Figure 3 Consistency among three statistical tests

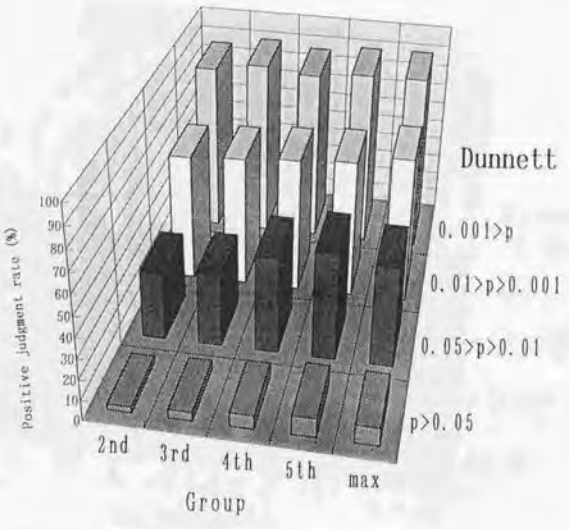


Figure 4 Consistency of change at each dose

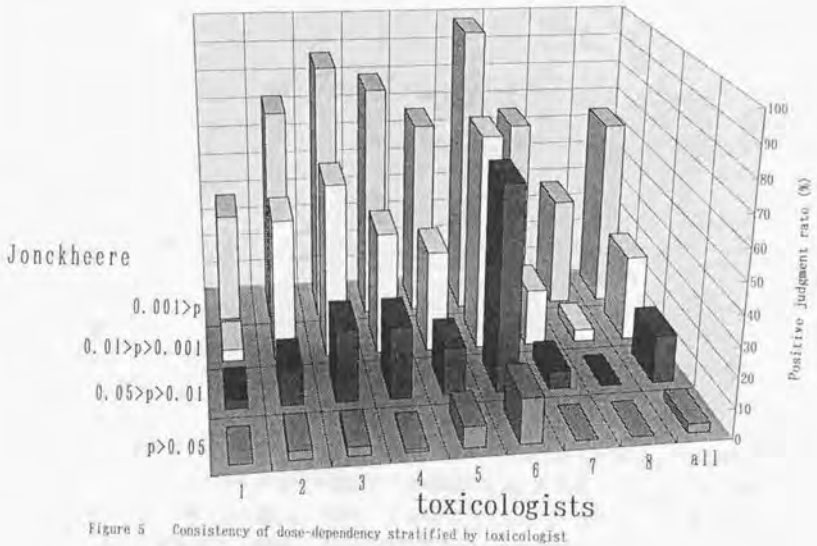


Figure 5 Consistency of dose-dependency stratified by toxicologist

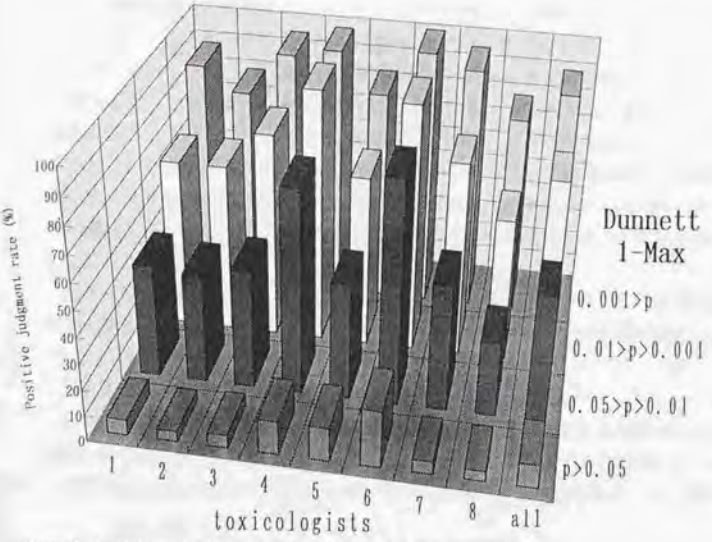


Figure 6 Consistency on change at maximum dose stratified by toxicologist

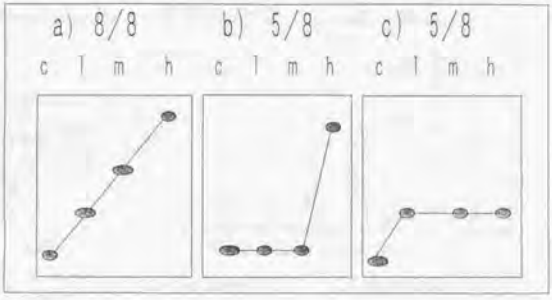


Figure 7 Difference of recognition for dose-dependency

Table 8 に，毒性家の用量相関性についての判断と最高用量での変化の関連を，評価者ごとに示した．最高用量の変化が陰性の場合に用量相関的な変化と判断している例は，全評価者を併せても，7項目・人とほとんどなく，用量相関的な変化であるためには，最高用量における変化は必要条件といえる．一方最高用量における変化が陽性でありながら，用量反応関係なしと判断する割合は，評価者によってかなり大きくばらついた．例えば2，3，6の毒性家は最高用量で変化がある場合には，ほとんど用量相関性ありと判断しており，最高用量での変化と用量相関性に対する判断の一致度が高い，これに対し他の評価者は，最高用量での変化が，用量相関性ありと判断するための十分条件とはなっていない．この原因について更に詳細に検討したところ，次の事実が明らかになった．Figure 7 に示したように，単調な用量反応関係として，いくつかのパターンを考えることができるが，何人かの毒性家は，このうちのいくつかのパターンを用量相関的な変化とは考えない．例えば，8人の毒性家全てが，直線的な変化であるパターン a)を用量相関的な変化と考えるが，1, 7, 8の毒性家は，最高用量でのみ変化があるパターン b)，最低用量から変化が生じその用量で反応が飽和に達するパターン c)を，用量相関的な変化とは判断しない．これらが連続的な変化でないためである．この3人の毒性家は，2段階以上の変化がみられる場合のみ，用量相関性ありと判断する．もちろん用量相関性の検定の検出力もこれらのパターンに依存して異なるが，その違いは量的なものであり，ある程度の大きさの変化があれば，統計学的にはみな有意になる．

用量相関の判定が項目グループ（血液・生化学・臓器重量）間ごとに異なるか，グループ別の評価も試みたが，グループ間で顕著な違いはなかったため結果は省略した．

・統計手法間で結果が食い違うケース

Table 9 に回帰分析とヨンキー検定の結果の一致度を示した．対角線上にあるセルの度数がかなり多く，2段階以上結果が異なるのは計12件に過ぎず，全般的には2つの方法で一致度は高いといえるが，頻度は少ないものの結果が大きく異なるケースも存在する．

Table 9 Consistency between parametric and nonparametric methods

Regression	Jonckheere				
	p>0.05	0.05>p>0.01	0.01>p>0.001	0.001>p	Total
p>0.05	1275	41	4	0	1320
0.05>p>0.01	44	124	26	5	199
0.01>p>0.001	2	33	95	21	151
0.001>p	1	0	13	317	331
Total	1322	198	138	343	2001

1）回帰分析（$p<0.01$），ヨンキー検定（$p>0.05$）(Figure 8)

このようなケースが計3件存在した．いずれも同じ試験のもので6群から構成される．最高用量群で大きく値が増加しているが，その手前の群まではわずかにメデアンが低下傾向にある．ノンパラメトリック手法の多くはデータの順位情報のみを利用するため，実スケールでの差の大きさを直接反映できない．この例では最高用量群における大きな変化が順位変換することで薄められるため，ヨンキー検定では有意となっていない．

2）回帰分析（$p>0.05$），ヨンキー検定（$p<0.01$）(Figure 9)

このようなケースが計4件存在した．共通点は外れ値が存在する点である．パラメトリックな手法は外れ値の影響を受けやすく，一般的にいえば外れ値が存在すると，検出力が低くなる．

このように2つの方法はそれぞれ一長一短がある．ノンパラメトリック手法についていえば，外れ値に対してロバストな点が長所とされているが，裏を返せば，毒性に起因するような異常値があったとしても，その影響を低めに評価してしまう．また実スケールでの大きな変化が反映できないという欠点を持つ．

Ⅳ．毒性家の判断と統計解析の一致度についての考察

4．1　外れ値の検出

・毒性家の外れ値に対する認識

毒性家は，外れ値を判定するにあたって，分布形とともにバラツキの大きさを考慮している．群内に相対的に外れた値があった場合でも，外れ方の大きさの絶対値が小さければ，外れ値とは判断されない．

毒性家は，これらの点に加え外れ値の生物学的異常性，他の観測値との関連についても考慮している．この点が，studentized residual で明らかに統計学的な有意性($p<0.00001$)が示されている場合であっても，毒性家の陽性判定率が，それほど高くなかった理由であると考えられる．例え一元配置型として，同一のデータが得られたとしても，項目の種類，他の測定値の値に依存して，毒性家の判断は異なる．これらの点を考慮して，背景データに基づいて構成した正常範囲との比較，マハラノビス距離などによる多変量的な外れ値の検討を行えば，より毒性家の判定に近づくと考えられるが，これらは通常の単変量の統計手法の限界を超えたものであり，今後の検討課題といえる．

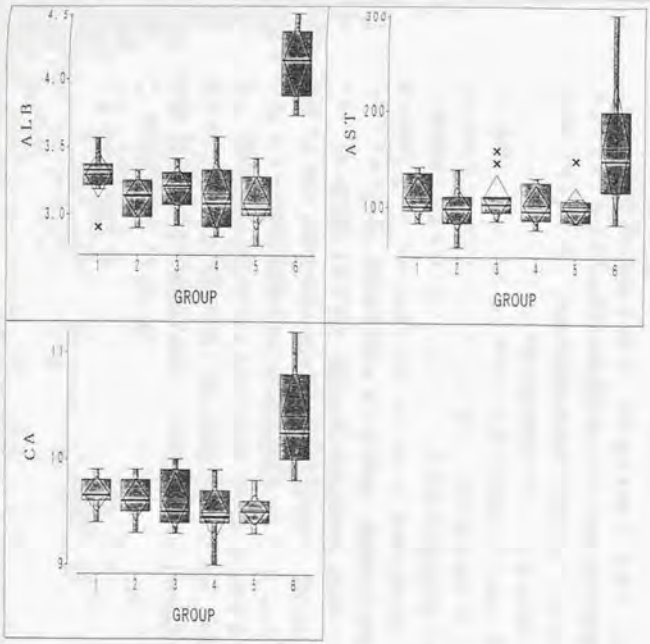


Figure 8 Boxplots with regression($p<0.01$) and Jonckheere($p>0.05$)
ALB:albumin (g/dl) AST: aspartate aminotransferase(IU/l) CA:calcium(mg/dl)
Diamond shape stands for mean ± SD

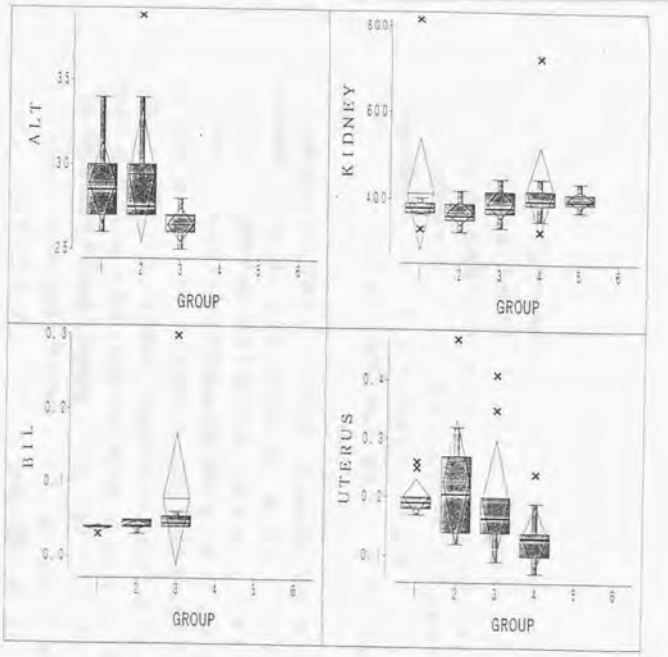


Figure 9 Boxplots with regression($p>0.05$) and Jonckheere($p<0.01$)
ALT: alanine aminotransferase(IU/l) BIL:total bilirubin(mg/dl)
KIDNEY:relative weight of kidney(mg%) UTERUS:relative weight of uterus(g%)
Diamond shape stands for mean ± SD

・外れ値を検出するための統計手法

統計学の役割は，生物学的な意味での異常値の候補をスクリーニングすることである．このためには，層別箱ひげ図，散布図などのグラフィカルな表現はたいへん有効ではあるが，1つの試験で100項目以上も評価しなければならない点を考えると，ルーティン的にこれらの図を描き，外れ値の有無を確認することは困難である．見逃しを防ぐためにも統計手法によって，外れ値を定量的に検出した上で，生物学的な意義を考察することが重要である．

外れ値を検出する方法として，studentized residualはskewnessとkurtosisより毒性家の判断に近くなった．studentized residualは，群間でプールした分散を用いるため，より安定した信頼性の高い結果が得られることになる．これに対しskewnessとkurtosisを計算するためにはそれぞれの群毎の分散を用いるため，不安定な結果を与える．

外れ値を検出するための他の方法として，箱ひげ図等を利用することも考えられるが，群毎にバラツキの大きさを評価する方法なので，skewnessとkurtosisと同様に結果が不安定になる．結果については示してないが，外れ値の検出について，箱ひげ図とkurtosisの性能が類似したものであることを確認している．またプールした分散を用いる統計手法として，studentized residual以外の手法を考えることもできるが，studentized residualは，外れ方の程度を直接表現しているので，結果が解釈し易いという利点を持っている．例えばstudentized residualの値がkであることは，外れ値がその値自身を除いた群平均から，k×SD離れた位置にあることを意味する．

studentized residualを用いて，外れ値を判断する目安としては，動物数の総数を50匹と考えて，Bonferroni法による調整を行うと，両側1%の棄却限界値は3.71となる．毒性試験は項目数が多く，あまり頻繁に外れ値と判定されてしまうと実質的な意味が失われることになるが，このように有意水準を設定すると，正規分布にしたがう100項目のデータに適用した場合，外れ値と判断される数の期待値は1であり，第1種の過誤をかなり小さく抑えることができる．このように厳しく有意水準を設定しても，毒性家の陽性判定率は20%程度とあまり高くなく（Table 3），異常値が見逃されるエラーはそれほど大きくはならない．Table 10にstudentized residualのcutoff点を3,4,5と変化させ，Table 3のデータについて，毒性家の判断に対する特異度，感度および擬陽性，擬陰性の確率を示した．cutoff点を3にすると，感度は85.41%と高いが，擬陽性が8.97%と高く，一般毒性試験では非常に多くの項目について統計解析を行うので，毒性家が陽性とは考えない場合でも，かなりの割合で統計学的には外れ値が生じてしまう．これに対しcutoff点を5にすると，擬陽性は0.68%とかなり小さくなるが，感度が40.86%と低下してしまう．cutoff点を4に設定すると，擬陽性は2.14%，感度は61.17%となり，感度と特異度がバランスされる．以上の結果からcutoff点として4前後が望ましいといえる．

Table 10 Specificity and sensitivity of studentized residual
Number of item·groups (row percent)

Judgment of toxicologists	Absolute value of studentized residual		
	SR<3	SR>3	total
negative	29271 (91.03%)	2886 (8.97%)	32945
positive	115 (14.59%)	673 (85.41%)	788
	SR<4	SR>4	total
negative	31468 (97.86%)	689 (2.14%)	32945
positive	306 (38.83%)	482 (61.17%)	788
	SR<5	SR>5	total
negative	31939 (99.32%)	218 (0.68%)	32945
positive	466 (59.14%)	322 (40.86%)	788

・外れ値のデータ解析の中での扱い

外れ値の原因としては，測定ミス，化合物の作用，化合物の効果とは独立な個体に起因する異常などが考えられるが，いずれの原因にせよ，外れ値は統計手法の選択に影響を与え，結果の解釈を困難にする．外れ値を棄却するための方法として，いくつかの統計手法，例えばSmirnov-Grubbs検定（吉村・大橋(1992)）が，標準的な統計学の教科書に記述されている．しかしながら統計学的な理由のみで外れ値を棄却すべきではない．なぜなら棄却検定の多くは，正規分布を前提としているため，有意になった場合，２つの可能性が考えられる．１つは観測値そのものの異常であり，もう１つは正規分布の仮定の違反である．多くのデータについて，厳密には正規分布を仮定することはできず，したがって棄却検定で有意になった場合でも，必ずしも測定値自体が異常であることを意味するものではない．

統計学的な手法によって，外れ値を検出した場合，まずその原因についてよく探索する必要がある．その原因が測定ミスであれば，可能であれば再測定を行い，可能でなければ，原因を明記した上で，その値を除去した上でデータ解析を行う．原因が測定ミスであることが明白でない場合は，基本的にはその値を含めてデータ解析を行う必要がある．特に毒性試験では，外れ値が生物学的にみて異常な値である場合は，値の存在を明記し，その原因が当該薬物に起因したものであるかについて考察する必要がある．また外れ値によって有意差が隠されている可能性があるので，異常値を除いたときに結果が変化するかを確認する感度分析を行う必要がある．外れ値が，生物学的にみて異常とは考えられない場合には，データが本質的に外れ値が出やすい正規分布以外の分布にしたがうことを意味している．正規分布を前提とした手法より，正規分布を前提としないノンパラメトリック手法を用いる，あるいはデータを変数変換して正規分布に近づけてから解析する必要がある．

・パラメトリック手法とノンパラメトリック手法の使い分け

この２つのアプローチは分布の形状が正規分布に近いか否かによって使い分ける必要があるが，通常の毒性試験のように１群の例数が10以下である場合には，分布の形状についての情報を得ることは困難である．この程度の規模のデータでは，正規分布にしたがうこ

とは，外れ値が存在しないこととほぼ同値であり，また3．1節で示したように，外れ値が存在すると2つのアプローチで結果が大きく食い違う．外れ値の扱いが，2種類のアプローチの使い分けを考える上で最重要な問題になる．ノンパラメトリック手法は外れ値の存在に対して結果が影響を受けにくく，歪んだ分布に対しても検出力が高いが，この点は毒性試験のデータを評価する上では必ずしもメリットとはいえない．毒性試験の重要な目的は，異常な値を示す個体の有無を調べることであり，ノンパラメトリック手法では異常値の影響を過小に評価してしまうためである．またノンパラメトリック手法は，順位情報のみを利用するため，Figure 8に示したように，実スケールにおける大きな変化が反映できない場合もある．このように2つのアプローチの使い分けは単純ではない．1つの対応策として，解析するデータについて予備検定を行い，どちらのアプローチを用いるかをデータ依存的に決定するツリー型アルゴリズムが長年用いられてきたが，Table 9で示したように2つのアプローチで結果が食い違うケースはそれほど多くなく，ツリー法によって性能が劇的に改善するわけではない．またツリー法では，同じ項目でも，性別あるいは時点によって解析方法が異なってしまうケースが生じ，解析手法の違いは結果の比較を困難なものにする．毒性試験は，薬物の安全性を確認するための探索的な側面を持ち，外れ値自体が毒性の可能性を示唆するので，外れ値の意味を考えずに，機械的にツリー型アルゴリズムを適用するのは好ましくない．

以下パラ・ノンパラ手法の使い分けについて述べる．

試験プロトコールを作成している段階で，測定項目を計量的なデータであるか，計数的なデータであるか分類しておく必要がある．計数的データについては選択すべき方法はノンパラメトリック手法であり，逆に計量データについては基本的にはパラメトリック手法を選択すべきである．外れ値が統計学的な手法によって検出された場合は，その理由をよく調べ，外れ値を除いた場合に結果が変わるかを確認しておくことが推奨される．外れ値を含めた場合と除いた場合の双方で，同様の結果が得られれば，結果の信頼性はより高まることになる．食い違う場合については，その理由を説明し，生物学的にどちらの結果の方がより最もらしいかを，考察する必要がある．このような感度分析を行うと，αエラーのインフレーションが起き，厳密には有意水準が保たれない可能性があるが，外れ値の原因が，化合物の毒性，個体の異常，実験手技のミス等であるとことを考えると，外れ値が存在する場合，革新的な立場で解析を行うことは，合理性をもっているといえる．

4．2 用量相関性についての解析

・用量相関性についての毒性家の認識

各用量における変化の評価については，統計手法と毒性家の判定の一致度は比較的高く，特にp値が0.001未満では毒性家は90%の確率で陽性と判定したのと対照的に，用量相関性については，統計手法と毒性家の判断の一致度はそれ程高いものではなかった．この原因は用量反応関係についての認識が，毒性家間で定性的に異なっていたためである．この結

果はわずか８人の毒性家に基づくものであり，結論の一般化可能性には当然限界があるが，今回評価を行った毒性家は，それぞれの研究所で指導的な立場にあり，毒性試験の現場では，用量相関性という用語の使い方に，混乱を起こしているといえる．

“用量相関性”あるいは“用量依存性”は毒性試験データを評価する上で，非常に重要な概念でありながら，明確な定義がないのが大きな問題である．用語自体が曖昧で，解釈の仕方に幅が生じている．学会活動などを通じてのコンセンサス作りが必要であるし，定義が不明確であるうちは，誤解をさけるため，直線に近い変化，単調的な変化などの，より正確な用語を用いる方が望ましいといえる．

・用量相関性の検定の有意水準

5%水準は人間が偶然かどうかを判断する基準に直観的にあっていると言われているが，用量相関性の検定の有意水準を5%とした場合，毒性家の判断と比べて，革新的で有意に出過ぎる傾向にあった．より厳しい有意水準である1%の方が毒性家の判断に近くなるといえる．毒性試験では用量相関性のみならず，対照群と各用量群の比較についても興味がある．特にこの研究では，毒性家はこの２種類の仮説について同時に評価した．したがって毒性家は，用量相関性と対照群との比較の２種類の仮説の多重性を無意識のうちに考慮し，厳しめに評価した可能性がある．

・用量反応パターンの解析

Figure 7に示したように，単調性のある用量反応関係としてはいくつかのパターンを考えることができるが，前述のように，一部の毒性家はb), c)のパターンを用量相関的な変化であるとは認識しなかった．ヨンキー検定などの通常の用量相関性の検定は，ある程度変化が大きくなるといずれも有意となり，検定の結果だけではこれらのパターンを区別することはできないが，いずれのパターンかによって，どのようなメカニズムで有害事象が発生したかの考察は当然異なってくる．毒性家は単に単調な用量反応関係を検出するだけでなく，用量反応関係が，S字型なのか，途中で飽和するのか，域値があるようなタイプであるかを明らかにできる解析手法を必要としている．次節では用量反応パターンを解析する方法として，最大対比法を示す．

V．最大対比法による用量反応パターンの解析

S字状の形状をとる用量反応曲線のモデルとしては，プロビット曲線，ロジスティック曲線などが知られており，培養細胞等を用いた in vitro 試験あるいは動物を用いた前臨床試験において，LD50 および ED50 を推定する背景のモデルとして用いられている．このようなモデルの下で閾値や飽和点を検討することも可能である．しかしながら一般毒性試験は，対照群を含め４ないし５群程度で実施されることが多く，あまり複雑なモデルを適用するメリットは少ない．このような状況で用量反応関係について検討する場合，用量反応

関係を複数の対比(contrast)によってモデル化する最大対比法を適用することが考えられる．最大対比法の概念を医薬研究の中で応用しようとする動きは，1990年頃から始まった．Ruberg(1989)はminimum effective doseを求めるために最大対比法を適用した．これとは独立に，大橋，浜田等は後期II相臨床試験において，用量反応関係を明らかにするため，最大対比法を用い始めた(浜田 他(1993) 浜田・岸本・大塚(1993))．また吉村，浜田等は毒性試験において，用量反応パターンを調べるために最大対比法を利用することを提案した(吉村・浜田(1996))．以下この方法について例解する．

一元配置分散分析型のデータ構造を想定し，Y_{ij}を第i群のj番目の観測値を表すものとする．　このとき帰無仮説の下で分散が1になるように基準化した対比統計量（Z）は，次のようになる．

$$Z = \Sigma C_i Y_{i\cdot} / \mathrm{SQRT}\, \Sigma [C_i^2 s^2 / n_i] \qquad (4)$$

ここで　C_i　：第i群の対比の係数　（Zの期待値を0にするため$\Sigma C_i = 0$とする）

$Y_{i\cdot}$　：第i群の平均値

n_i　：第i群のサンプルサイズ

s^2　：誤差分散　$\Sigma\Sigma [Y_{ij} - Y_{i\cdot}]^2 / \Sigma [n_i - 1]$

毒性試験では設定する用量範囲がかなり広いため，様々な用量反応関係が生じる可能性があり，それぞれの用量反応パターンを検出するため，Table 11に示した対比を設定することが考えられる．(d)のexpの対比は，少し特殊であるが，毒性試験では，0, 1, 3, 10というように3倍公比で等比級数的に用量が設定されることが多く，ここでは反応の大きさが指数的に増加する場合を検出するために，この対比を加えている．

Table 11 Coefficients of maximum contrast method

	control	low-dose	mid-dose	high-dose	
(a)	−3	−1	1	3	linear
(b)	−5	−1	3	3	m-end(saturated at middle dose)
(c)	−3	1	1	1	l-end(saturated at low dose)
(d)	−7	−5	−1	13	exp(like exponential curve)
(e)	−3	−3	1	5	m-start(middle threshold type)
(f)	−1	−1	−1	3	h-start(high threshold type)

また本論文では，対比のあてはまりを評価するために，対比の群間平方和に対する寄与率をモデル適合率(MF：Model Fitness)として特に定義することにする．モデル適合率は次のように計算される．

$$MF = SS / \Sigma n_i [Y_{i\cdot} - Y_{\cdot\cdot}]^2 \qquad (5)$$

ここで$SS = \Sigma [Y_{i\cdot} C_i]^2 / \Sigma [C_i^2 / n_i]$である．

MFは0〜1の範囲の値をとり，値が高い程，対比の群間平方和に対する説明力が高いことを意味する．MFが1のときは，その対比によって群間平方和が完全に説明される．このように複数の対比統計量の値を同時に計算し，その最大値Z_{max}に基づいて，用量反応パターンについての情報を得るのが最大対比法である．用量反応関係を評価する際，偶然を越えた意味ある用量反応パターンのみを拾い上げる必要がある．このためには検定が有効な手

段であるが，複数の対比について検定を行うと，多重性の問題によって，第一種の過誤が増大するため，有意水準を調整した解析が必要になる．任意の複数の対比に対して多重性を調整する古典的な多重比較法として，Bonferroni，Scheffe 法が知られているが，どちらも過度に保守的になる．特にこの例のように対比統計量間の相関が高いときは，非常に保守的になってしまう．岸本，浜田は，任意の複数の対比について，等分散性と多変量正規分布を前提として，多重積分によって多重性調整ｐ値を計算するプログラムを作成しているが（岸本・浜田(1994)），計算に時間がかかり，指定する対比の種類，サンプルサイズによって棄却限界値が異なるため，汎用的なアプローチとはいえない．

これに対し，多重性に関する調整ｐ値を resampling（標本再抽出法）によって計算することを，Westfall and Young(1992)は提案している．この方法は正確な並べ替え検定のｐ値を近似するものである．また SAS MULTTEST はこの目的のために開発されたプログラムである(SAS Institute(1996) 浜田・吉田(1992) 浜田(1996))．

標本再抽出法では，次の手順によって多重性調整ｐ値を計算する．

1）観測データについて，複数の対比のそれぞれについて検定を行いｐ値を計算する．

2）観測データから非復元無作為抽出を行って（群番号を並べ替えて），元の観測データと同じサンプルサイズの擬似的な標本を作成する．

3）擬似的な標本に対し１）と同じ検定を行いｐ値を計算する．このうち最小のｐ値をＭｉｎ　ｐ値とする．

4）２），３）を繰り返して，Ｍｉｎ　ｐ値の分布関数Ｆを十分な精度で推定する．この繰り返し数を標本再抽出回数とよぶ．

5）１）の個々の検定について生のｐ値がＭｉｎ　ｐ値より小さくなった回数を，再抽出回数で割ったものが調整ｐ値となる．

・最大対比法の適用例

４群で行われた毒性試験の赤血球データ（Table 12）に基づいて，最大対比法の適用例を示す．このデータでは薬物投与によって赤血球数の減少傾向が観測されている．

Table 12　RBC(red blood cell counts) data　(unit：$\times 10^4/mm^3$)

Group　Dose	Raw　data($\times 10^4/mm^3$)	Mean	SD
Control　0mg	925 917 912 912 949 908 908 989 931 909	926.0	25.7
low-dose　1mg	898 925 908 873 908 941 893 920 922 931	911.9	20.1
mid-dose　3mg	874 876 916 908 873 807 874 919 952 916	891.5	39.8
high-dose 10mg	869 919 874 852 830 906 914 898 933 935	893.0	35.4

Table 13　Result of maximum contrast method

Contrast	non-adjusted p-value	adjusted p-value	Z	MF (model fitness)
linear	0.0104*	0.0258*	-2.7052	0.876
m-end	0.0067*	0.0177*	-2.8778	0.992
l-end	0.0224*	0.0532	-2.3868	0.682
exp	0.0426*	0.1018	-2.1017	0.529
m-start	0.0216*	0.0515	-2.4013	0.695
h-start	0.1491	0.3249	-1.4742	0.260

*　:significant at 5% level

散布図をFigure 10に示した.

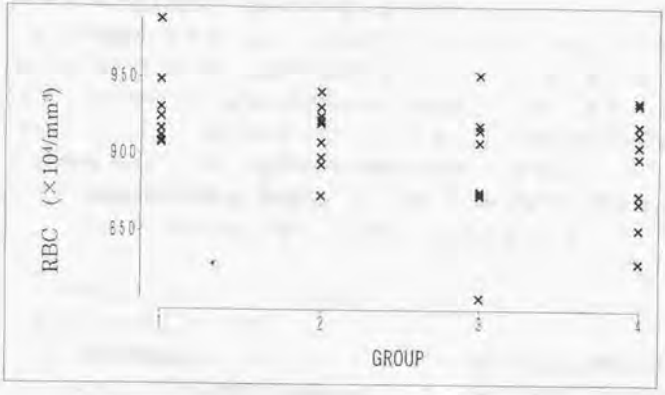


Figure 10　Scatter plot of　RBC　(×10⁴/mm³)

6本の対比をあてはめた結果をTable 13に示した. 表中でnon-adjusted p-valueは多重性について調整せずに, 自由度36のt分布で近似した両側p値, adjusted p-valueはSASのMULTTESTプロシジャによって10000回の標本再抽出によって調整した両側p値である. またZとMFは, それぞれ対比統計量と, 群間平方和に対する対比のモデル適合率を示している. 調整前のp値では5本の対比が, 両側5%の水準で有意であるが, 調整したp値で有意となるのはlinearとm-endの2つのみである. Z統計量の絶対値が最大の対比はm-endで, MFが0.99であるので, 群間平方和のほぼ全てをm-endという対比によって説明できることがわかる. すなわち中用量までは直線的に赤血球数は低下し, 中用量で飽和に達することがわかる. linearの対比についても有意ではあるが, モデル適合率は0.90以下でm-endよりかなり劣っている.

VI. 最大対比法の性能評価

以下ではシミュレーション実験により最大対比法の検出力を調べ, 回帰分析と比較する

（6．1節）．また最大対比法による用量反応パターンの解析結果と，人による判断との一致度について評価する（6．2節）．

6．1　シミュレーション実験による最大対比法の検出力の評価

対照群を含めて4群の場合について，Table 11に示した6種類の対比を設定した場合の最大対比法の検出力を評価した．ここで検出力は，対立仮説の下でいずれかの有意な用量反応関係を検出できる確率として定義した．

次に示すシミュレーション条件において，最大対比法と，回帰分析（-3, -1, 1, 3の対比を用いた場合），対立仮説に合せて理論的に最も検出力が高くなるように対比を選んだ場合（以下では最適法と呼ぶ）の検出力を比較した．以下のシミュレーションでは簡便のため，等分散（分散既知）と正規分布で，かつ例数のバランスがとれていることを前提としたが，比較する手法は全て等分散と正規性を仮定したパラメトリック手法であるため，この条件が成り立たない場合でも，性能が相対的に大きく異なることはないと考えられる．

標本再抽出法によって個々のシミュレーション実験ごとに，多重性調整p値を計算するのは時間がかかるので，まず正規分布と等分散性を前提に，100万回のシミュレーション実験によって両側5%点を求めた．帰無仮説の下でZ統計量の最大値の棄却限界値は2.36となった．これは個々の対比を有意水準0.018で検定することに相当する．

・シミュレーション条件

1）群構成：4群（対照群＋3用量群）

2）サンプルサイズ：1群あたり10

3）4群の期待値をμ_1，μ_2，μ_3，μ_4として，SASのrannnor関数を用いて，$N(\mu_i, 1)$正規乱数を発生させた．μ_iについてはTable 14に示した．

4）シミュレーション回数：それぞれの期待値の組み合わせについて100000回

5）有意水準:両側5%（棄却限界値：2.36(最大対比法)　1.96(回帰分析，最適法)）

6）適用手法

最大対比法，回帰分析，最適法（Table 14に対比の係数を示した）

7）対立仮説：Table 14に示した10通りの対立仮説を想定した1～6までは最大対比法で想定した対比に対応する用量反応関係，7は最大対比法では想定してない中用量で階段状に変化するパターン，8-10は非単調な用量反応関係である．

・シミュレーションの結果

Table 15にシミュレーションによって求めた検出力を示した．

1）linear(1)，m-end(4)，m-start(5)の3つの用量反応関係については，回帰分析の方が最大対比法に比べ検出力が勝るが，その違いは大きくても5%程度である．また最大対比法の検出力は，最適な方法と比べても悪い場合で7%下がる程度である．

Table 14 Alternative hypotheses

No	Content of D-R relationship	Name	μ_1, μ_2, μ_3, μ_4	coefficients of optimal contrast
1	linear	linear	Δ(-1, -1/3, 1/3, 1)	-3, -1, 1, 3
2	exponential	exp	Δ(-1, -0.8, -0.4, 1)	-7, -5, -1, 13
3	saturated at low-dose	l-end	Δ(-1, 1, 1, 1)	-3, 1, 1, 1
4	saturated at Mid-dose	m-end	Δ(-1, 0, 1, 1)	-5, -1, 3, 3
5	middle-threshold	m-start	Δ(-1, -1, 0, 1)	-3, -3, 1, 5
6	high-threshold	h-start	Δ(-1, -1, -1, 1)	-1, -1, -1, 3
7	middle-threshold saturated at Mid-dose	l-m	Δ(-1, -1, 1, 1)	-1, -1, 1, 1
8	quadratic	quadratic	Δ(-1, 1, 1, -1)	-1, 1, 1, -1
9	downturn	downturn1	Δ(-1, 0, 1, 0)	-1, 0, 1, 0
10	downturn	downturn2	Δ(-1, 0, 1, 0.5)	-9, -1, 7, 3

Δ: 0 ～ 1 by 0.1

2) exp(2)の用量反応関係については，最大対比法と回帰分析では検出力はほとんど変わらない．最適な方法と比べて，この２つの手法の検出力は5%程度劣る場合がある．

3) l-end(3)とh-start(6)の用量反応関係については，回帰分析と比べて最大対比法の方が検出力が高く，最大10%以上の差が生じる．また最適な方法と比べた場合でも，最大対比法は10%以上劣ることはない．

4) 低用量まで変化がなく，中用量で飽和する用量反応関係l-m(7)については，最大対比法を構成する６本の対比には，対応するものを含めていないため，この場合には回帰分析の方が検出力が高くなるが，その違いはせいぜい5%程度である．最適な方法と比べても，検出力の違いは大きくても10%強である．

5) ２次曲線型の用量反応関係があるときには，回帰分析の検出力はΔが増えても全く増加しないが，最大対比法では，最適法と比べて劣るものの，ある程度の検出力は保持できる．

6) 頭打ちがあるとき downturn(9, 10)には，最大対比法の方が回帰分析より検出力が高く，特に downturn1(9)では，20%以上高くなる場合もある．

以上の結果をまとめると，単調性のある様々な用量反応関係について，最大対比法と回帰分析で検出しやすいパターンに多少の違いはあるものの，全体的にはその差は小さく，最大対比法の検出力は，回帰分析と比べて同等であるといえる．また，最適な対比を１本のみ用いる場合と比べてもそれほど性能が落ちるものではない．２次曲線的な変化や，頭打ち現象などが起こり，単調性が成り立たない場合については，最大対比法の方が回帰分析より検出力がかなり高くなる．

Table 15 Results of simulation

			Δ										
No.	D-R shape	Method	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1	linear	MCM	0.049	0.072	0.138	0.261	0.420	0.599	0.759	0.877	0.947	0.982	0.994
1	linear	REG	0.050	0.077	0.155	0.293	0.469	0.655	0.806	0.910	0.964	0.989	0.997
1	linear *	OPT	0.050	0.077	0.155	0.293	0.469	0.655	0.806	0.910	0.964	0.989	0.997
2	exp	MCM	0.049	0.072	0.142	0.269	0.440	0.628	0.788	0.901	0.961	0.988	0.997
2	exp	REG	0.050	0.075	0.147	0.274	0.440	0.619	0.773	0.886	0.950	0.983	0.995
2	exp	OPT	0.051	0.079	0.165	0.317	0.505	0.696	0.840	0.933	0.977	0.994	0.999
3	l-end	MCM	0.049	0.073	0.149	0.293	0.484	0.689	0.846	0.941	0.982	0.996	0.999
3	l-end	REG	0.050	0.072	0.135	0.247	0.395	0.565	0.719	0.844	0.924	0.970	0.988
3	l-end	OPT	0.049	0.084	0.195	0.376	0.591	0.780	0.907	0.971	0.992	0.999	1.000
4	m-end	MCM	0.049	0.075	0.152	0.295	0.478	0.674	0.830	0.929	0.976	0.994	0.999
4	m-end	REG	0.050	0.080	0.166	0.318	0.506	0.697	0.842	0.933	0.977	0.994	0.999
4	m-end	OPT	0.050	0.083	0.181	0.349	0.551	0.746	0.880	0.957	0.987	0.997	1.000
5	m-start	MCM	0.049	0.075	0.153	0.295	0.480	0.676	0.831	0.929	0.976	0.994	0.999
5	m-start	REG	0.050	0.080	0.166	0.318	0.506	0.697	0.842	0.933	0.977	0.994	0.999
5	m-start	OPT	0.051	0.083	0.181	0.350	0.553	0.747	0.881	0.955	0.987	0.997	0.999
6	h-start	MCM	0.049	0.074	0.149	0.294	0.489	0.691	0.846	0.941	0.982	0.996	0.999
6	h-start	REG	0.050	0.072	0.135	0.247	0.395	0.565	0.719	0.844	0.924	0.970	0.988
6	h-start	OPT	0.050	0.086	0.195	0.375	0.591	0.783	0.906	0.969	0.992	0.998	1.000
7	l-m	MCM	0.049	0.081	0.178	0.358	0.579	0.785	0.913	0.974	0.994	0.999	1.000
7	l-m	REG	0.050	0.089	0.203	0.396	0.617	0.808	0.923	0.977	0.995	0.999	1.000
7	l-m	OPT	0.050	0.098	0.241	0.475	0.714	0.885	0.966	0.993	0.999	1.000	1.000
8	quadratic	MCM	0.049	0.065	0.117	0.214	0.361	0.548	0.732	0.877	0.955	0.988	0.998
8	quadratic	REG	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.051	0.049	0.050
8	quadratic	OPT	0.050	0.096	0.245	0.474	0.714	0.887	0.966	0.994	0.999	1.000	1.000
9	downturn1	MCM	0.049	0.061	0.096	0.165	0.266	0.397	0.545	0.690	0.811	0.898	0.949
9	downturn1	REG	0.050	0.059	0.087	0.135	0.204	0.292	0.395	0.507	0.620	0.722	0.806
9	downturn1	OPT	0.049	0.073	0.145	0.269	0.428	0.609	0.765	0.880	0.948	0.980	0.994
10	downturn2	MCM	0.049	0.067	0.119	0.217	0.357	0.524	0.691	0.826	0.917	0.967	0.989
10	downturn2	REG	0.050	0.068	0.121	0.215	0.343	0.493	0.644	0.775	0.874	0.940	0.973
10	downturn2	OPT	0.049	0.076	0.153	0.288	0.460	0.647	0.799	0.906	0.963	0.988	0.997

MCM: Maximum contrast method REG: regression analysis(-3, -1, 1, 3) OPT:optimal method

*:The result is same as that of REG. Because optimal contrast is -3, -1, 1, 3 in this case.

6. 2 最大対比法による解析と人による判断の一致度の評価

最大対比法による用量反応パターンの判定と毒性家の判断の一致度を評価した.

・方法

1) 毒性学の専門家7人と, 統計解析の業務従事者5人をノミネートした.

2) 研究用データベースから, 4群で行われた試験のデータを選択した. 4)で示す a) ~f)

の用量反応関係を各２通り（モデル適合率が 0.95 以上と 0.90 前後になるものを選択した），計 12 通り，g)のタイプとして，中用量から立ち上がって，中用量で飽和するタイプを２通り，非単調な場合を２通り，h)として有意な用量反応関係がない場合を２通り，合計 18 項目を選択した．

3）18 項目について Table 11 に示した６本の対比統計量を計算し，各対比の調整しない p 値，標本再抽出法（抽出回数 10000 回）によって多重性について調整した p 値，モデル適合率(MP)を計算した．

4）１）の 12 人が独立に，無作為に順序を並べ替えた 2)の 18 項目の，生データ，要約統計量(平均値，SD)，散布図に基づき，a)～h)の８種類の用量反応関係に分類した(付録２参照)．

a)直線的に増加（等差的に増加）　b)指数的に増加（等比的に増加）
c)低用量で飽和　d)中用量で飽和
e)中用量から立ち上がり　f)高用量から立ち上がり
g)　a)～f)以外の用量反応関係がある　h)用量反応関係は認められない

・結果

Table 16 に 1-18 の用量反応関係の人による判断と最大対比法による判定が比較できる形で，結果をまとめてみた．この表では同一の項目で複数のデータがある場合，BW，BW2 のように数字を付けて区別している．また Figure 11 に用量反応関係の散布図を示した．12 人の評価者は７人の毒性家と５人の統計解析の業務従事者から構成されているが，用量反応パターンの判定について専門分野間で顕著な差がみられず，層別すると人数が減ってしまうため，12 人をまとめた結果のみを以下では示す．最大対比法によって a)～f)と判定される用量反応関係が１～12 まで示されており，13, 14 は最大対比法では想定していない，対比の係数を-1, -1, 1, 1 にとったとき大きな値をとる用量反応関係である．13 と 14 については参考として，下段に-1, 1, 1, 1 の対比のモデル適合率が示されている．15, 16 は低用量で上がって，中用量で下がって，高用量で再び上がる非単調な用量反応関係がある場合である．このようなケースでは最大対比は有意となるものの，モデル適合率は 0.50 前後と低く，最大対比法で想定した用量反応関係とは異なることがわかる．17, 18 は用量反応関係が認められないケースで，最大対比の調整 p 値が 0.05 より大きくなる．

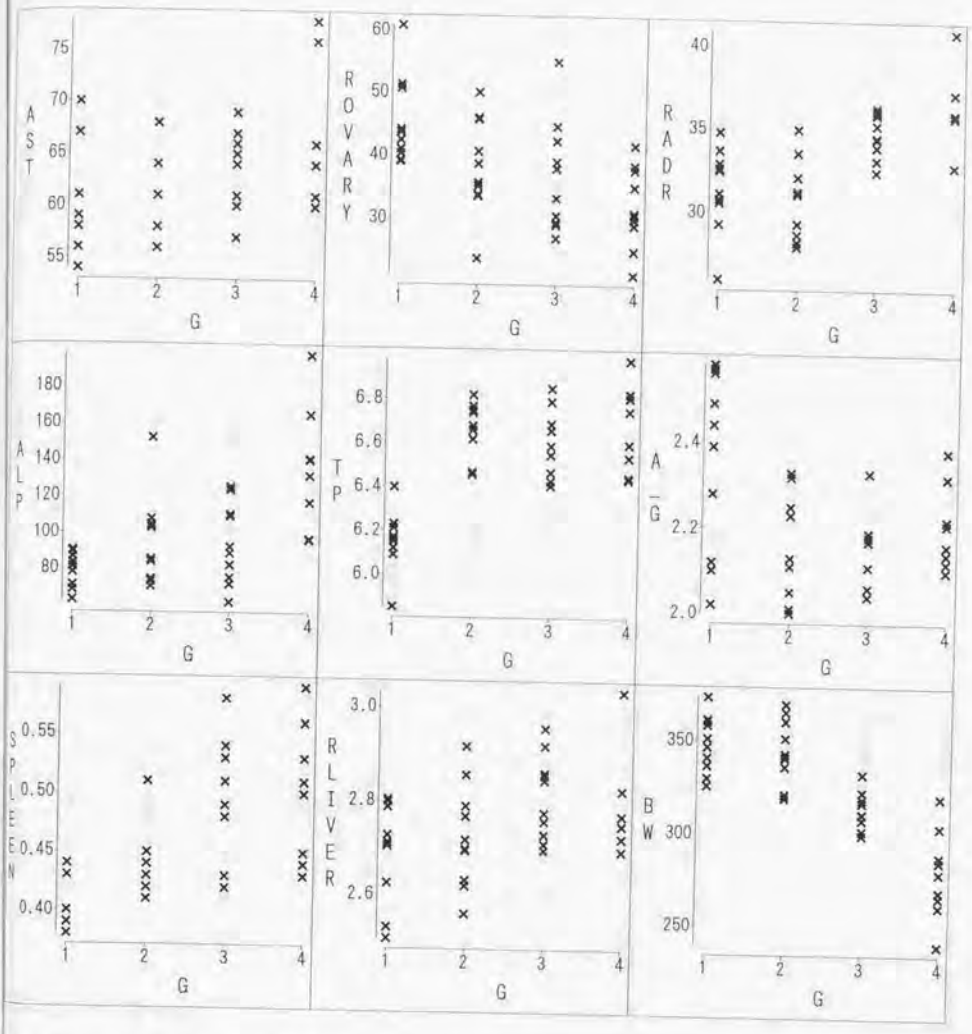


Figure 11 Scatter plots of sample data

AST:aspartate aminotransferase ROVARY:relative ovary weight RADR2:relative adrenal weight
ALP:alkaline phosphatase TP:total protein A/G:albumin globulin ratio SPLEEN:spleen weight
RLIVER:relative liver weight BW:body weight THYMUS:thymus weight
RKIDNEY:relative kidney weight BW2:body weight RHEART:relative heart weight
RADR:relative adrenal weight TC:total cholesterol RPIT:relative pituitary weight
TG:triglyceride PLT:platelet count

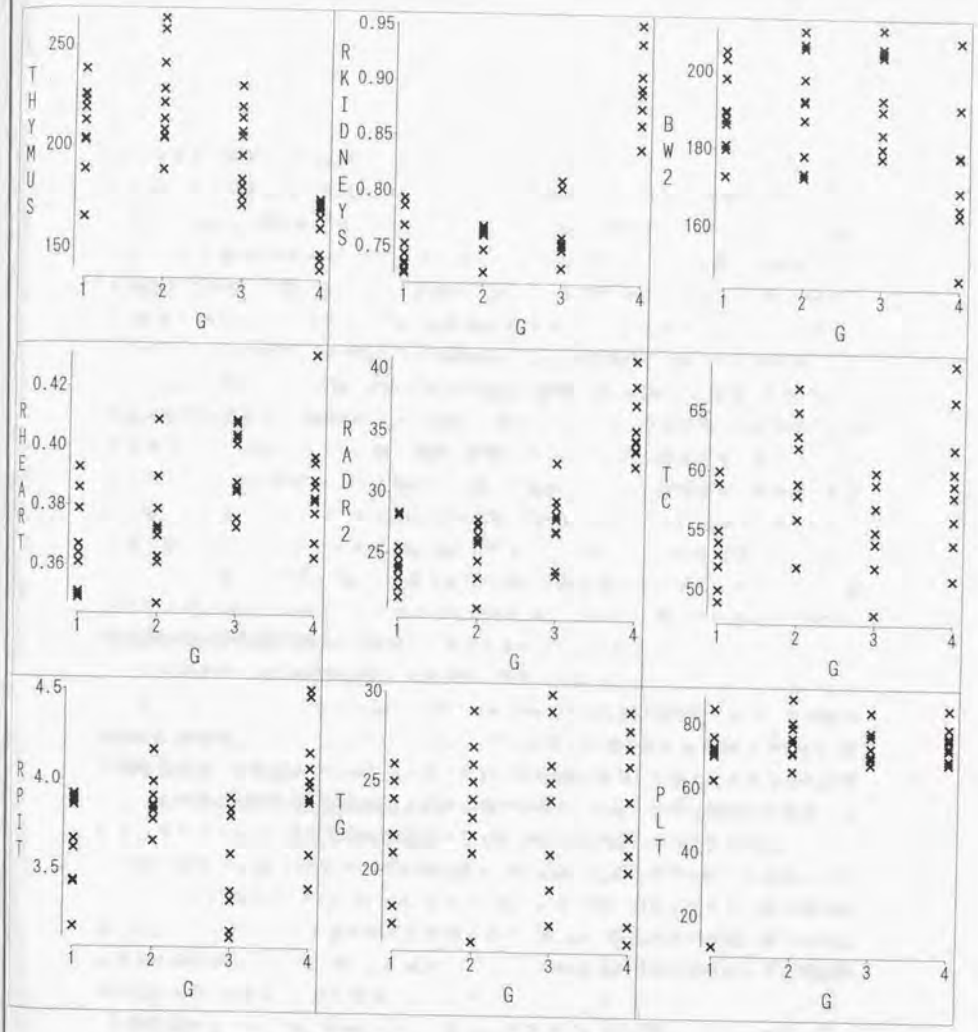


Figure 11 continued

1）用量反応関係がa)～f)のケース

評価者間で判定に大きなばらつきがあることがわかる．12人の評価者間で結果が完全に一致するのは12項目中2項目(ROVARYとTP)のみである．残りの10項目の中で，6番のA_Gについては12人中9人までは，低用量で飽和(c)，残りの3人は用量反応関係なし(h)と判断しており，用量反応パターン自体は1通りにしか判断されてない．11番のRKIDNEYと12番のBW2についても12人中11人までは，高用量から立ち上がるf(h-start)と判断しており，異なる判断をしたのは1人のみであり，ほぼ評価者間で一致した結果が得られている．残りの項目については，例えばASTのように調整p値(0.0000)が高度に有意でかつ，対比のモデル適合率(0.9997)が0.90を超える場合であっても，評価者間で判定が異なる場合がある．この原因については後で個別に説明する．このように評価者間で判定にバラツキがあるものの，7番のSPLEENを除いて，最大対比法の判定は，評価者間の最多数の判定に一致している．SPLEENについても，最大対比d(m-end)に対し，a(linear)の判定が6人と最も多くなっているが，5人はd(m-end)と判定している．1～12の12項目については，それぞれの用量反応関係に対応した最大対比の調整p値は5%水準で有意であり，モデル適合率も0.90を越えているので，この2つの条件が満たされれば，最大対比法による判定は，ほぼ評価者の最多数の判定に一致すると考えてよいことがわかる．

以下評価者間で判定が分れたケースを順次，説明していく．

1番のASTについては，a(linear)6人と，f(h-start)3人に判定が分れている．平均値はきれいに直線的に上がっているためlinearの対比のモデル適合率は0.9997と高いが，散布図をみると，中用量までは対照群の最大値以下の範囲で分布しており，大きな変化はないが，高用量において外れ値気味に2個体の値が上昇している．この2個体のみ変化していると考えた3人は，高用量のみ変化があるf(h-start)と判定したと考えられる．

3番のRADRと9番のBWについてはb(exp)とe(m-start)に判定が分かれているが，この2つは対比の係数が-7，-5，-1，13・-3，-3，1，5と似ているため，区別しにくかったためである．RADRではexpのモデル適合率が0.9998に対し，m-startのそれは0.9305，BWではexpのモデル適合率が0.9407に対し，m-startのそれは0.9978と，両方の対比のモデル適合率がともに0.90を越え，大きな値をとっている．

4番のALPについては，b(exp)5人，f(h-start)3人，g（その他）3人，a（linear）1人の4通りに判定が分れている．4群のそれぞれの平均値は80，96，94，134となり，低用量と中用量とで平均値が逆転し，平均値は用量に対して単調には変化してない．この原因として，低用量群に高い方で外れた値が存在し，平均値を引っ張っていることがあげられる．この外れた値以外では，用量の増加に伴い，指数的に反応が上昇している．外れた値の影響を割り引いて評価した人は，b(exp)と判定し，平均値の値を重視した人は，高用量のみ変化するf(h-start)に判定したと考えられる．

7番のSPLEENについては，a(linear)6人とd(m-end)5人に判定がほぼ2つに割れているが，これは，この2つの対比のモデル適合率が高く，似通った値を取っているためであ

る．モデル適合率はd(m-end)が0.9978，a(linear)が0.9407となっている．

8番のRLIVERについてはd(m-end)7人，a(linear)2人，e(l-end)1人，f(h-start)1人，g(その他)1人の5通り，10番のTHYMUSについてはe(m-start)7人，g(その他)4人，f(h-start)1人の3通りに判定が割れている．各群の平均値はRLIVERでは2.69，2.72，2.81，2.80，THYMUSについては211，224，199，162となり，どちらも単調性が成り立っていない．RLIVERについてはほぼ単調性が成り立っているような印象も受けるが，高用量群に値の高い外れ値が存在し，平均値を引っ張っており，この値を除くと頭打ちのような現象がみられる．THYMUSについては低用量で一旦値が上昇してから下降する傾向が，グラフから読み取れ，これが4人もg(その他)と判定した原因である．

調整p値が有意であり，モデル適合率が0.90を超えるときは最大対比法による判定は，評価者間の最多数の判定に一致するといえるが，外れ値が存在する，平均値が用量に対して単調に変化してない，似たようなモデル適合率の対比が複数存在するケースでは，評価者間で判定が分れる場合がある．

2）用量反応関係がgのケース

13番のRHEARTと14番のRADR2は，最大対比法を構成する6本の対比では含まれていない-1，-1，1，1の対比が最も大きくなる場合である．このようなケースでも，最大対比法でモデル適合率が最大となる対比に対応する用量反応関係が1番多く選択されている(RHEART:d(m-end)6人，RADR2:e(m-start)8人)．しかしながらg(その他)の判定も，それぞれ5人と3人と多く，a)～f)以外の用量反応関係があると判断した人も多かったことが判る．中用量で階段状に変化する用量反応関係を検出することに関心が高ければ，-1，-1，1，1の対比を最大対比法の構成対比に加える必要があるが，通常用量反応関係は連続的であり，このような用量反応関係はかなり用量範囲を広くとらない限りは，あまり現実的とはいえない．

非単調な用量反応関係がある15番のTCと16番のRPITの場合では，最大対比は5%水準で有意になるもののモデル適合率は低い．このようなケースで最大対比に対応する用量反応関係を選択する評価者は12人中1人のみであり，残りの評価者は，h（用量反応関係なし）か，g（その他）と判定している．

3）用量反応関係がhのケース

有意な用量反応関係がない17番のTGと18番のPLTでは，最大対比に対応する用量反応関係e(l-end)を選択する評価者も，それぞれ，1人，2人存在するものの，残りはh（用量反応関係なし）を選択しており，最大対比法で有意とならない結果を反映したものといえる．

以上，最大対比法で考慮した用量反応関係があるケース，想定しない用量反応関係があるケース，用量反応関係なしのいずれの場合においても，評価者間で多少の判定のバラツ

キはあるものの，最大対比法による用量パターンの解析結果は，評価者間で最も多い判定に一致し，人間の直観に合致したものといえる．

Table 16 Judgment for dose-response patterns

No	Item	Content of item	Maximum contrast	Raw p value	Adjusted p value	MF	Judgment by researchers
1	AST a)	aspartate amino-transferase	a(linear)	0.0024	0.0071**	0.9997	a(6)f(3)g(1)b(1)h(1)
2	ROVARY a)	relative ovary w.	a(linear)	0.0004	0.0015**	0.9443	a(12)
3	RADR b)	r. adrenal w.	b(reg)	0.0000	0.0000**	0.9998	b(9)e(3)
4	ALP b)	alkaline phosphatase	b(reg)	0.0000	0.0000**	0.9445	b(5)f(3)g(3)a(1)
5	TP c)	total protein	c(l-end)	0.0000	0.0000**	0.9923	c(12)
6	A_G c)	albumin globulin ratio	c(l-end)	0.0015	0.0051**	0.9188	c(9)h(3)
7	SPLEEN d)	spleen w.	d(m-end)	0.0000	0.0000**	0.9893	a(6)d(5)e(1)
8	RLIVER d)	r. liver w.	d(m-end)	0.0092	0.0246*	0.9181	d(7)a(2)c(1)f(1)g(1)
9	BW e)	body weight	e(m-start)	0.0000	0.0000**	0.9978	e(8)b(4)
10	THYMUS e)	thymus w.	e(m-start)	0.0000	0.0000**	0.9305	e(7)g(4)f(1)
11	RKIDNEY f)	r. kidney w.	f(h-start)	0.0000	0.0000**	0.9429	f(11)e(1)
12	BW2 f)	body weight	f(h-start)	0.0034	0.0114*	0.9020	f(11)h(1)
13	RHEART g)	relative heart w.	d(m-end) -1 -1 1 1	0.0006 0.0005	0.0013**	0.9040 0.9178	d(6)g(5)h(1)
14	RADR2 g)	r. adrenal w.	e(m-start) -1 1 1 1	0.0000 0.0000	0.0001**	0.8572 0.9881	e(8)g(3)a(1)
15	TC g)	total cholesterol	c(l-end)	0.0050	0.0143*	0.4536	h(8)g(3)c(1)
16	RPIT g)	r. pituitary w.	f(h-start)	0.0030	0.0094**	0.5136	h(6)g(5)f(1)
17	TG h)	triglyceride	c(l-end)	0.3718	0.6784	0.8772	h(11)c(1)
18	PLT h)	platelet count	c(l-end)	0.0389	0.0673	0.9719	h(10)c(2)

( ) : the number of person

*: 5% significant **:1% significant

Ⅶ. 最大対比法についての考察

最大対比法は，回帰分析と比べて同等以上の検出力を持っており，また調整p値が有意かつモデル適合率が0.90以上であるときは，最も多くの人が判断する用量パターンに一致することがわかった．以上の結果は，用量反応パターンを統計学的に客観的に判断する手段として，最大対比法の妥当性を示すものといえる．以下計算のテクニカルな面を含めて，最大対比法の問題点について考察する．

・標本の抽出回数

標本再抽出法によって最大対比法の調整p値を計算する際，標本の再抽出回数をどのように設定するかが問題となる．標本再抽出回数は多い方がp値の推定精度は高くなり，正確な並べ替え検定のp値に近づくが，その見返りとして再抽出回数に比例して計算時間も増大する．p値の誤差分散の大きさは，二項分布の分散の公式によりp(1-p)/Nとなる．ここでpはp値であり，Nは標本の再抽出回数である．p値は5%前後のときの精度が最も問題となるので，p=0.05として95%の信頼区間を求めてみた．

N	推定標準誤差	95%信頼区間(p=0.05)
100	0.0218	0.0064～0.0936
1000	0.0069	0.0362～0.0638
10000	0.0022	0.0456～0.0544
100000	0.0007	0.0486～0.0514

標本の再抽出回数が1万回になれば，p値の95%信頼区間は0.05±0.0044%となり，1%の範囲に収まるので，実質的に十分な精度がある．標本再抽出回数が1万回程度であれば，最新のコンピュータハードウエア，ソフトウエアを用いれば，十分実用的な時間内に計算でき，標本再抽出法の利用の大きな妨げとはならない．例えば著者の所有するPentium(200MHz)をCPUとして用いたパーソナルコンピュータで，MULTTESTプロシジャを使用して，Table 12のデータを解析するのに必要な時間は数秒である．

・他の用量相関性の検定との性能比較

本研究ではTable 11に示した6本の対比を用いた場合について，最大対比法の検出力を回帰分析と比較し，同等以上の性能があることを示したが，用量相関性の検定としては，他にMax t検定（栗木・広津・Hayter(1989)），Williams法(Williams (1971))などが知られており，これらの統計手法も回帰分析と同等以上の検出力を有している（栗木・広津・Hayter(1989)）．以下これらの手法と最大対比法の関連について述べる．Max t検定，Williams法を対比を用いて説明すると，

MAX t法は，4群の場合，

-3, 1, 1, 1　　-1, -1, 1, 1　　-1, -1, -1, 3

という3種類の対比のうちの最大値

Williams法は，最高用量と対照群を比較する場合

-1, 0, 0, 1　　-2, 0, 1, 1　　-3, 1, 1, 1

の3種類の対比の最大値に基づいて検定を構成することにほぼ相当する．したがって広義には両手法とも，最大対比法に含めてよい手法であり，対比の選び方が異なるのみである(Yoshimura, Wakana and Hamada(1997))．

MAX t法のうち-3, 1, 1, 1と-1, -1, -1, 3，Williams法のうち-3, 1, 1, 1の対比は，Table 10の6本の対比の中に含まれている．この点から類推されるように，構成される対比によっ

て，検出し易いパターンは異なるものの，この3手法間で統計学的な性能，特に検出力は大きく異ならない．MAX t 法と Williams 法は元来，それぞれ階段上の変化，対照群と比べて変化を起こしている用量を検出するために導かれた方法であり，その結果から用量反応関係についての情報を直接得るのは困難である．これに対し提案法では，直接用量反応関係と対応するように対比を選んでいるので，結果からどのような用量反応パターンがあるかを判断でき，解釈し易く，アクションに結びつき易いという利点がある．また本研究で示したように，その判定結果は，人による判断に一致する．このように最大対比法は，用量反応パターンを客観的に判断する新しい統計手法として位置付けることができる．ただし最大対比法を用いる場合，単に対比の有意性の評価のみでは，用量反応パターンを判断する情報としては十分ではなく，モデル適合率についても検討する必要がある．

・対比の選択

本研究では Table 11 の6本の対比を設定したが，exp と m_start の対比は係数が似ているため，対比統計量間の相関がかなり高くなる．このためこの2つの用量反応関係は区別しにくくなっている．exp の対比を除いても，両側5%の棄却限界値は 2.36 から 2.35 とほとんど変わらず，統計学的な性能自体はほとんど変化しない．用量が等比的に設定されていない，あるいは用量に反応が直接比例するような用量反応関係に関心が薄ければ，この対比を除いても大きな問題はない．

また目的に応じて，対比の選択の仕方には工夫が必要である．例えば本研究では毒性試験の問題を取り上げたが，臨床試験の第II相で用量反応関係を調べる場合，用量の範囲は動物を用いた毒性試験と比べるとかなり狭く設定され，興味の対象はどの用量で反応が飽和するかにある．このような場合，Table 17 のように最大対比法の構成対比を設定するのが合理的である(岸本・浜田(1994))．

Table 17 Coefficients of the maximum contrast method for phase II clinical trials

control	low-dose	mid-dose	high-dose	
−3	−1	1	3	linear
−3	1	1	1	l-end
−5	−1	3	3	m-end

本研究では4群の場合で最大対比法を定式化し，その性能を評価したが，3群や5群で行われた試験についても，用量反応パターンを検討したい場合がある．このようなケースでは，4群の場合を修飾して最大対比法を使用する必要があるが，検出したい用量反応パターンを対比によってモデル化し，その最大値によって用量反応パターンを判定する最大対比法の考え方自身は，異なるものではない．また5群の場合についても，最大対比法の検出力は，回帰分析と比べて同等以上であることが報告されている(上松・三宅(1997))．

Ⅷ. 結論と今後の課題

本研究では，毒性学の専門家の判断と統計解析の結果を比較することにより，望ましい毒性試験の統計解析手法について検討した．その結果，以下の点が明らかになった．

1）外れ値を検出する手段としては studentized residual が毒性の専門家の判断に近くなり，推奨できる．

2）用量相関性についての認識が，毒性家間で質的に異なることが判明した．

3）用量相関性の検定の有意水準としては，通常用いられる 5%より厳しい 1%程度の水準の方が毒性家の判断に近くなる．

4）用量反応のパターンを評価する方法としての最大対比法の利用を提案し，その性能を評価した．回帰分析と比べて，最大対比法は同等以上の検出力を持ち，かつそのパターン判定は，人による最多数の判定に一致した．用量反応パターンを客観的に評価する方法として，最大対比法の妥当性が示されたといえる．

今回の研究では小動物を用いて行われる一般毒性試験の計量データを対象としたが，サンプルサイズが 1 群 3-5 匹で行われる大動物の試験，あるいは奇形の発生率等のカテゴリカルデータが多く生じる生殖試験等でも，同様の研究を行い，標準的な統計手法を推奨していく予定である．なお本論文で得られた結果については，ＤＩＡ，毒科学会，製薬協，医薬安全性研究会等で既に紹介しているが，その成果を毒性試験の現場で，より有効に利用してもらうため，教育活動，あるいは提案方法のコンピュータプログラムを作成し，現場への普及を進めていく予定である．

Ⅸ. 謝辞

本研究に参加していただき，お忙しい中，毒性試験のデータおよび用量反応パターンの判定をしていただいた毒性学の専門家の，日本たばこ産業株式会社　松本一彦氏，第一製薬株式会社　野村護氏，ファイザー製薬株式会社　飯島護丈氏，日本農薬株式会社　永見俊之氏，日本化薬株式会社　半田淳氏，サンド薬品株式会社　斎藤実氏，大鵬薬品工業株式会社　佐野正樹氏，万有製薬株式会社　池田孝則氏の方々に心から感謝申し上げます．

また用量反応パターンを判定していただいた統計解析の業務従事者の，SAS Institute Japan　岸本淳司氏，三井製薬株式会社　平河威氏，三菱化学株式会社　酒井弘憲氏，第一製薬株会社　森田智視氏，ヘキストジャパン株式会社 清見文明氏の方々および，集計を手伝っていただいた日本たばこ産業株式会社 吉野 慶氏，三菱化学株式会社　阿部いくみ氏にも心からお礼申し上げます．

さらに研究をご指導頂きました東京大学医学部薬剤疫学講座助教授　久保田潔先生，東京理科大学工学部経営工学科教授　吉村功先生，また貴重なご意見をいただきました東京大学医学部疫学生物統計学講座助教授　橋本修二先生，東京大学医学部薬剤疫学講座助手矢船明史先生，同じく東京大学医学部薬剤疫学講座助手　小出大介先生に甚大なる感謝の

意を表します.

X. 参考文献

1)厚生省薬務局審査第一課(1991) 医薬品毒性試験法ｶﾞｲﾄﾞﾗｲﾝ1990 解説. 薬事日報社

2)山崎実, 野口雄次, 丹田勝, 新谷茂(1981) ラット一般毒性試験における統計的手法の検討 対照群との多重比較のためのアルゴリズム. 武田研究所報 40, 3, 163-187

3)K. Kobayashi, K. Watanabe and H. Inoue(1995) Questioning the usefulness of the non-parametric analysis of quantitative data into ranked data in toxicology studies. The Journal of Toxicological Sciences, 20, 47-53

4)浜田知久馬, 岸本淳司(1995) ＳＡＳによるノンパラメトリック多重比較. 計算機統計学, 8-1, 77-83

5)SAS Institute Inc. (1996) SAS/STAT Software:Changes and Enhancements through Release 6.11. SAS Institute Inc.

6)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

7)L. A. Hothorn, K. K. Lin, C. Hamada and W. Rebel(1997) Recommendation for biostatistics of repeated toxicity studies. Drug Information Association Journal, 31-2, 327-334

8)吉村功(1995) 統計学は良薬を作るのに役立つか. 数理科学, 389, 43-499)

9)竹内啓監修(1989) 統計学辞典. 東洋経済新報社

10)S. J. Ruberg(1989) Contrasts for identifying the minimum effective dose. Journal of the American Statistical Association, 84, 816-822

11)浜田知久馬, 岸本淳司, 大塚芳正, 大橋靖雄(1993) 後期第Ⅱ相試験における指摘用量設定のための統計学的アプローチ. 第61回日本統計学会講演報告集, 56-57

12)吉村功, 浜田知久馬(1996) 医薬データの統計解析における最大対比法の活用. 第64回統計学会予稿集, 42-43

13)浜田知久馬, 岸本淳司, 大塚芳正(1993) 臨床試験における至適用量設定のための統計学的アプローチ. 応用統計学会予稿集, 40-44

14)岸本淳司, 浜田知久馬(1994) 任意の対比群について多重比較を行う数値積分プログラム. 計算機統計学, 7-2, 147-154

15)P. H. Westfall and S. S. Young(1992) Resampling-Based Multiple Testing. John Wiley and Sons

16)浜田知久馬, 吉田道弘(1992) MULTTEST プロシジャの紹介. SUGI-J'92 論文集, 357-370

17)浜田知久馬(1996) ＳＡＳによる用量相関性の解析. SUGI-J'96 論文集, 331-346

18)栗木哲, 広津千尋, A. J. Hayter(1989) 累積カイ２乗の最大成分に基づく多重比較法 ―有意確率と用量水準比較への応用―. 日本応用統計学会誌, 18, 129-141

19)D. A. Williams (1971) A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics, 27, 103-117

20) I. Yoshimura, A. Wakana and C. Hamada (1997) Performance comparison of maximum contrast methods to detect dose dependency. Drug Information Association Journal, 31-2, 423-432

21) 上松潤哉, 三宅慎太郎 (1997) 用量パターンの推定における最大対比法の性能評価. 東京理科大学 工学部第一部経営工学科 卒業研究論文

付録１　用量相関性と外れ値の評価シート

Study No.　　male　　Name

NO.	ITEM	Outlier				Dose-dependency	Change at each dose		
1	RBC	1	2	3	4		2	3	4
2	Hb	1	2	3	4		2	3	4
3	Ht	1	2	3	4		2	3	4
4	WBC	1	2	3	4		2	3	4
5	PL	1	2	3	4		2	3	4
6	GOT	1	2	3	4		2	3	4
7	GPT	1	2	3	4		2	3	4
8	ALP	1	2	3	4		2	3	4
9	CPK	1	2	3	4		2	3	4
10	tG	1	2	3	4		2	3	4
11	GLU	1	2	3	4		2	3	4
12	TCh	1	2	3	4		2	3	4
13	BUN	1	2	3	4		2	3	4
14	Cr	1	2	3	4		2	3	4
15	TP	1	2	3	4		2	3	4
16	Alb	1	2	3	4		2	3	4
17	A/G	1	2	3	4		2	3	4
18	TBi	1	2	3	4		2	3	4
19	Na	1	2	3	4		2	3	4
20	K	1	2	3	4		2	3	4
21	Cl	1	2	3	4		2	3	4

毒性試験データに判定に関するお願い

１．各群ごとに、異常値、外れ値があるかどうかを判定して下さい．

２．各項目ごとに、用量相関性について評価してみて下さい

３．各項目ごとに、どの用量に変化が起きているかを示して下さい．変化が起きていると思われる用量には、すべて丸を付けて下さい．

付録２ 用量反応パターンの評価シート

所属＿＿＿＿＿＿＿＿ 氏名＿＿＿＿＿＿＿＿

どのような用量反応関係があるか判定してください。

1）ラットを用いた４群の一般毒性試験で計 18 項目あります。各測定項目は同じ個体のものではなく独立なデータであることに注意してください。用量は 0, 1, 3, 10 のようにほぼ等比的に設定されていると考えてください。

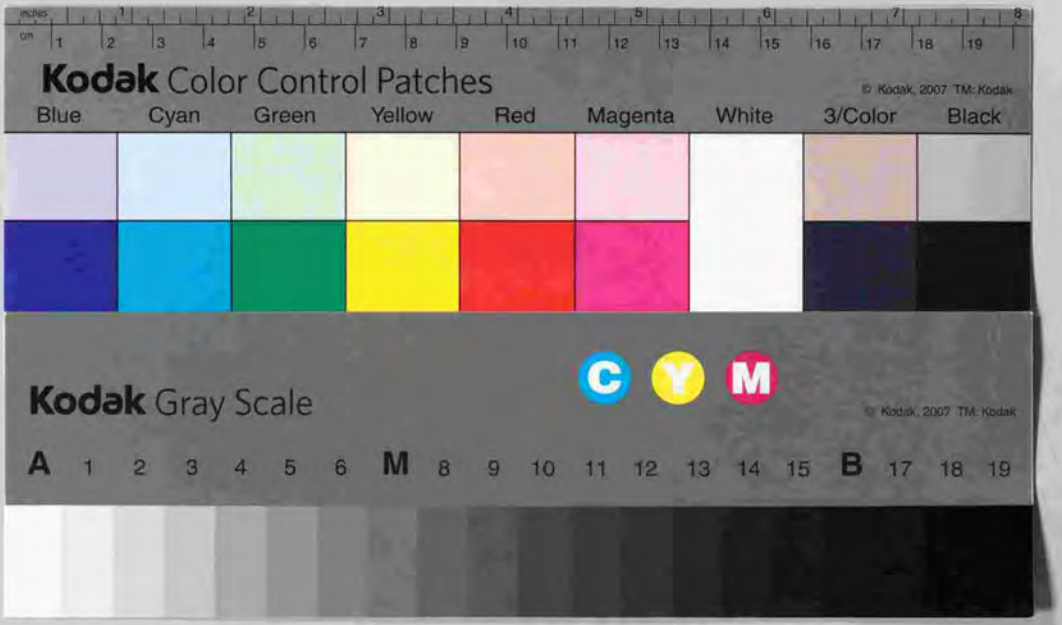
2）判定材料として生データ、平均・ＳＤ、グラフを用意しました。

3）用量反応関係については

a）～h）に分類してください。またｇ）に分類する場合は、必ずどのような用量反応関係であるかコメントしてください。その他何か気づいた点があればコメントしていただいて結構です。

a）直線的に増加（等差的に増加）
b）指数的に増加（等比的に増加）
c）低用量で飽和
d）中用量で飽和
e）中用量から立ち上がり
f）高用量から立ち上がり
g）a）～f）以外の用量反応関係がある
h）用量反応関係は認められない

番号、項目名	分類	コメント
① TP		
② ALP		
③ TC		
④ TG		
⑤ AST		
⑥ A/G		
⑦ PLT		
⑧ BW		
⑨ BW2		
⑩ SPLEEN		
⑪ THYMUS		
⑫ RHEART		
⑬ RLIVER		
⑭ RKIDNEY		
⑮ RPIT		
⑯ RADR		
⑰ RADR2		
⑱ ROVARY		


inches
cm
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
Kodak Gray Scale
C
Y
M
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19