ISSN 0288-2906

# 医薬安全性研究会

会 報 No.17
July 1985

目 次

# 長期および短期試験に於ける発癌性の評価*
## （抄　録）

訳；財）安評センター
小 林 克 己

現在多くの人々が実験動物を用いて実際に癌原性試験を行い、併せて癌発生率の把握の為統計処理を行っている。その結果、しばしば試験に対して適合性を欠いている。
第二次大戦後、癌原性試験を解釈する為、統計手法は大いに発展し、より単純化し、不偏性を持ち、真の発癌性をより繁栄している。
asymptotically efficient法は非常に簡単に理解でき、あまり統計に理解がなくても応用できる。これは本来投薬グループ動物の腫瘍発生を比較する場合異なった試験で実施した値と直接比較検討するのに補正を用いている。以来、改良材料のないまま実施されて来たが、今後これら腫瘍の統計手法は発癌テストに於いて標準化されなくてはならない。これら発癌テストはメーカー，政府，ブリーダの人達から、実際動物試験によって要求されている。

ＩＡＲＣ（国際対癌連合）によって腫瘍発生の統計処理法が文章によって公開される事を感じている著者らは、この腫瘍検定法が統計学者でない人、またアンチ統計学者にも説明したい。我々は統計よりもむしろ英語で書くことを試みた。これらの手法は誤解をさける為、異なった方法で多くの時間と多くの言葉遣いを費やし、読者をまどわさないように心がけた。我々は項目を細分化し、各見出しは分類し、大文字で何が書いてあるかを文章に示し分類してあるこの方法は、すでに本法を知っている者は飛ばすことが出来る。
Section 4 は統計Noteからなり、終わりまで記載されている。そして興味あるデータについて統計処理法を実施した。

しかしながら、この全文章を読んでさえ十分に思い出すことは出来ないであろうが、心配しないで下さい。君は実際分析する為に腫瘍データを持つまで待ってこのデータに対して推奨される方法でトライしなさい。
本文中はほとんどポケット計算機で対応でき、もし便利なコンピュータを持っていれば、我々は便利なプログラムを持っているのでリチャードにリクエストすれば手に入れられる。

30年前から古典的統計原理の非統計言語を総括した文章が多くある。
腫瘍の発生には３つの種類に分けられ、各々統計処理される。偶発的（incidental），致命的（fatal ），単純化（mortality-independent ）。上記の致命的と単純化とはむしろどちらかと云えば同意である。偶発的と致命的および単純化とは異なった手法を用いる。本法を理解するには、はじめの半分を飛ばして読み、Section 4 へ進むとよい。実際のデータを引用しSection 4.4 では行っている。そしてSection 5.3 と 5.4を特に読んでほしいと願っている。

---

* International Agency For Reseach On Cancer,
Monographs, supplement 2, LYON 1980

## 1．2 総　括

特定の物質が発癌性が有るかどうかの決定にはこの10年間に数千の試験によって導かれて来たが、このデータの中にはほとんど発癌性の判断が出来ない結果を含むデザインまたは実行がなされている。
しかしながら、これらのデータはむしろ説明の解釈が深く、要求に対して完璧に実行しているものも有る。
目だった相違点は処理グループ間で腫瘍発生が偶発か真実かである。
逆に偶発性は毒性によるもののどちらかでも、真の腫瘍の異なった解釈の原因ともなることがある。

動物の配分方法，飼育方法，剖検や病理検索などで片寄り的な問題が有るとすれば発癌試験の結果、腫瘍の発生数に実験群間の差が生じてくる理由としては次の３つの可能性が考えられる。
１．試験で生存している動物が癌になる危険の時間的差。
２．一見して同じに見える生存動物が癌になる予測困難なチャンスの差。
３．それぞれの群間で被験物質そのものの発癌的影響の差。
理論上、我々はまず第１の時間の長さを修正し、次いで第２のチャンスの差として考えるのは余りに差異の大きなものを第３の被験物質の影響によるものと考えていく必要がある。
従って、このようなデータを説明して行くに３つのステップが必要である。

First 段階
動物数は死亡によるもので、死亡動物は特定のタイプの腫瘍に分類され、全群と同様の発生条件として期待値を算出し比較する。
期待値は特定され、全試験群に於ける平均発生値可能値である。
腫瘍発生数（観察数）が期待値より低い群がいくつか有っても観察数（値）は必ず何らかの群を越えなくてはならない。そして最後に全期待値は和されて腫瘍発生を正確に検定する。

Second段階
高薬量群に於いて観察値に比べて期待値が大きい場合は片側検定を用いて確率を判定する。
対照群に比べてポジティブな傾向のみ検定を行う。

Third 段階
判定はいずれにせよP値で決められる。
水準は 1% Level で行う。
有意水準は必ずコンバイントしたものについて行うこと。

1) 本検定は他の試験と同一の要因をもつ。
2) 背景データの腫瘍発生にも応用可能。
3) 推定ターゲット臓器中の他の病変または初期癌かどうかについて他のテストまたは本テストから把握できる。
4) 種々の短期テストに於ける被験物質の影響あるいはその逆もわかる。
5) 構造相関には信頼できない。

本統計の分析の主なねらいはむしろ試験の重要性より発生数の大きな差について記載されている。我々は各処理間に於ける癌の発生状態をディスプレイ上検討したい。
もし死亡した動物に興味ある癌が発見されれば生存動物に対してもこのタイプを採用し分類をはじめる。Kaplan-Meier法、皮膚および体表の腫瘍で死亡したものは無関係とする。しかし、K-M 法では、死亡動物の腫瘍は分類されていない。確かなタイプの偶発的腫瘍の場合我々はそれなりの分類法で行う。これはMaximum Likelihoodと云う無関係に死んだ動物の偶発的な腫瘍に用いられている。Kaplan-MeierとMaximum Likelihoodは図式され、記述式の為計算が楽である。しかし、ルーチンワークでは用いられない。(差があったのみに行うのである。)

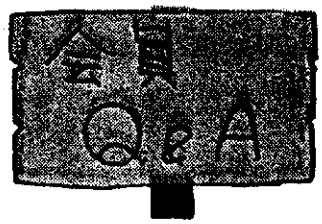
全データの評価の本質は腫瘍発生が生存時に有ったかどうか深く検討し偏見のない様にルーチ化する事。研究室内のルーチンと本報告とでは若干の相違がある。これは試験者によってコンティクスコードがルーチン化されていない為である。
1＝明確に偶発的
2＝たぶん偶発的
3＝おそらく致命的
4＝明確に致命的
5＝死亡に関係なし

本手法はポケット計算機で十分対応できる。しかしコンピュータは信頼でき便利である。データの細部について経時的に受領できる。これらのプログラムは小さく作ることが簡単である。本プログラムは簡単に用いられる為、また他のプログラムから楽に変換できる様にフォートランで書かれている。

発癌性をテストする際、ここに述べる方法よりもより教養のある統計的方法を使うことから得られる科学的利益はふつう何もない。そして実のところ、もし他の統計的方法がexperimentalists concernedあるいは聴衆者に理解しにくい方法に代わって使われるならばいくらか不利がある。
形式が単純で統計学者でない人にとって使いやすく理解しやすいという事実にもかかわらず、推められた統計的方法は、真の発癌性が癌の発生率をかえる見込みのある方法に対して、先入観がなく、感度の高いという多くの特性を持っている。
しかしこの文書は、統計学者でない人のためのもののつもりであるから、統計学者のみの興味の主なポイントのすべて（e.g.推められた方法が統計的に最適である理由）は、Section 5.4 の中で小さな型にまとめ（relegated ）られ、それはほとんどの読者に調べられる必要がない。

Section 1 ：要約の拡張と、Section 4 ：一般例（その中には長生きするための矯正の方法が明確にされている）との間にSection 2 ：背景的な考慮とSection 3 ：INTERCURRENT MORTALITY に注目されるグループの間に実質的（重要）な相違点がない特別な場合がくる。Section 2 とSection 3 は共に全く標準的な統計的考慮を検討している。そして、多くの読者に表面的にざっと目を通されるべきであろう。
同様に、付表１と２は省略され得るが、付表３（Section 5.3 ）は腫瘍関係の動物の率の一般的な使用や実験結果を述べるためのmean latencyによって生じるかもしれないいくつかの誤りについて討議しており、それは、一般的な興味のものである。最後に、付表４（Section 5.4 ）は、統計学者のためだけであるが、統計的な“効率”と推められた方法の“偏見のなさ”を討議している。（それぞれのサブセクションは、すでに知っている人が省略できるように１つの文かあるいは、大文字で書かれている２つのいわゆる要約が先に書かれている。）

腫瘍観察の、基本的に違う３つの背景は“附随的”“重大”“mortality-independent ”である。




# スコアデータにおける検出力の問題

住友製薬(株)長田明彦

**Q** 外用薬の抗炎症作用を2薬剤で比較する試験を行ないました。薬剤を塗布したモルモットの背部皮膚に紫外線を照射し、発赤の程度を採点しました(0~12点, 整理点)。

観測値からは正規分布を仮定しがたく思われたので, 平均値の差の検定をWilcoxonの順位和検定で行なったところ差はありません($P<0.05$)。

この場合、検出力はどのように考えれば良いでしょうか。

データ

| 薬剤 | 動物数 | 紅斑スコア | 平均スコア |
|---|---|---|---|
| A | 8 | 1 1 1 0 0 0 4 0 | 0.875 |
| B | 8 | 0 1 0 1 0 0 0 1 | 0.375 |

**A** (吉村先生)

(1) 検出力の求め方: 検出力を計算するためには, まず値の確率分布の想定が必要である。この例で度数分布は次のようになる。

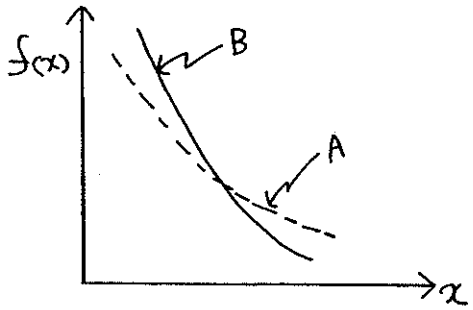
| 薬剤＼スコア | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| A | 4 | 3 | 0 | 0 | 1 |
| B | 5 | 3 | 0 | 0 | 0 |

これに合う分布は, 例えば指数分布とかポアソン分布のようなものである。指数分布で考えてみると

A薬剤: $f(x)=\theta_A e^{-\theta_A x}$

B薬剤: $f(x)=\theta_B e^{-\theta_B x}$

と表せる。




$\theta_A$ と $\theta_B$ がどれくらい違えば Wilcoxon 検定で検出できるかを調べれば良い。但し、数式的には出来ないので、コンピュータを使ったシミュレーションによる解析となる。AとBの分布にあうような乱数を8個ずつ発生させ、その有意性を Wilcoxon 法で検定することを10万回ほどやらせれば良い。大型のコンピュータでは数分で終わる。

(2) 留意すべきこと：この例題を考える場合、検出力の計算以前の問題がありそうである。データ中で、はずれ値は4点が1つだけである。これが2点として出てくればどうか（恐らくは、薬剤間の差について問題にならないだろう）。また、これが12点であれば果して気になるかどうかということになる。

Wilcoxonの順位和検定では、これが2点であろうが、12点であろうが統計量としては同じになる（いずれにしても最高値であるから）。つまりこの点数の大きさにどれ程の意味があるかが問題となる。もし2点と12点との発赤反応に明らかな差があって、スコアの平均値に意味ありとすれば、Wilcoxonの検定よりも点数の大きさに意味を持たせた別の方法を考えるほうが良いことになる。

いずれにしても、1つだけ"はずれ値"があってそれによって結論が決ってしまうというのは危険であり、一応の計算は行なうにしてもその解決には注意が必要である。

(3) Wilcoxonの検定について

Wilcoxonの順位和検定というのは、そもそも分布が等しくて、平均値だけがずれている場合の検定である。分散が等しくないときはそれを前提としたWelchの検定が良い。

Q 基礎講座で学んでいるDecision treeでは、まず等分散の検定(バートレット)を行ない、有意であればノンパラメトリック法に行くことになっているが、分散が違うからノンパラメトリックへ行くのは間違いということになるのでしょうか。

A Decision treeを形式的に適用するのは時により非常に変なことになり、注意が要る。まぁまぁの目安ぐらいに考えるのが良い。分散がかけ離れているのに平均値を比較して意味があるのかという問題もある。

(4) 観測者が複数いる場合

Q 観測者が2人いて、その合計を点数にすることに不都合はないでしょうか。

A 観測者が複数いる場合、その平均をとるとバラツキが減少するという意味で、問題はないと思われる。

Q 1匹のモルモットの6ヶ所に0～3点のスコアをつける方法において、観測者(複数)と測定場所についてすべて合計したものをスコアとしても良いものでしょうか。

A 場所の違いについては、別々のデータとしておいて、二元配置の分析で場所による差があるかどうかを調べ、差がないなら平均してしまって良い。

長谷氏(グレラン製薬)のコメント 観測者が複数の場合 その合計点をとると、両極端でのスコアの出現頻度は低くなるので、合計点の意味は絶対的なものを反映せず、むしろ相対的な比較に意味があるということだろう。従って、スコアの点数と具体的な変化の程度が対応するためには、1人ずつが独立してスコアをつけ、観測者を要因として二元配置の問題として考えるほうが良いのではないか。

参考文献

柳川堯："ノンパラメトリック法" 培風館 ―新統計学シリーズ

皆様からの質問お待ちしております。事務局

Statistics for Toxicologists
P309- P310

(訳) 日本農薬㈱
永見 俊之

## 体重と臓器重量

動物に化学物質を連続的に投与（暴露）する試験で，通常集めることができる一連のデータの中に体重と特定の臓器重量のデータがある。体重はしばしば悪影響が現われる最初のパラメータである。

体重データの最善の分析方法および臓器重量データ（絶対重量，重量変化，あるいは体重百分率として）の分析形式は，過去の多くの論文（41,72,73,77）の主題であった。

我々の経験上，サンプルサイズが十分大きい（10以上）場合，下記の１－４の手順が適切であった。サンプルサイズが小さい場合，データの正規性は不確かとなり， Kruskal-Wallisのようなノンパラメトリックな方法がより適切であろう。（ref 81参照）

1.臓器重量を体重の百分率として計算する。
2.体重は絶対重量か体重変化のいずれかにより分析できる。たとえ群を試験開始時に正しくランダマイズ（いずれの群も相互に平均体重に有意差がなく，全ての群の全動物の体重が総平均の標準偏差の２倍の巾以内にある。）したとしても，計算上やや厄介であるが体重変化を用いる方が有利である。
3.分散の均一性を確かめるため，各組のデータにBartlettの検定を施す。
4.適当な場合には，章の最初に示した決定樹に略述された一連の分析に従う。

## 臨床化学

多くの臨床化学のパラメータは，現在慢性毒性試験における動物から採取される血液や尿により測定される。過去においては（現在でもいくつかの機関で），１変量のパラメトリックな方法（主にｔ検定およびANOVA またはその一方）によるデータの評価が容認されてきた。しかし多くの理由によりこれが最善のアプローチではないことを示すことができる。

第一にそのような生化学的パラメータは，めったに互いに独立ではない。また我々の関心が，これらパラメータのただ一つに集中することもしばしばはない。むしろ一連のパラメータとして特定の標的臓器に対する毒作用と関連をもつ。

たとえばcreatine phosphokinase（CPK），α－hydroxybutyrate dehydrogenase（α－HBDH），乳酸脱水素酵素（LDH）の増加が共に起きることは心筋の障害を強く示唆する。そのような場合，我々はこれらのパラメータの単に一つに関心を示すのではなく，3つのパラメータの全てに関心をもつ。

表10に種々のパラメータと特定の標的臓器との関連について簡単な概略を示す。同様に血清電解質（ナトリウム，カリウム，カルシウム）の変化は相互に影響し合う。あるイオンの減少は他の一つのイオンの増加に結びついている。更にデータの性質として（あるパラメータの場合には），パラメータの生物学的背景のために，あるいはその測定方法のために正規分布（ガウス分布）していないか，もしくは本来連続的でない（たとえば，クレアチニン，ナトリウム，カリウム，塩素，カルシウムの場合には，Mitruka や Rawbskey の実験動物の引用データ参照）。これらの性質はこの章で述べたパラメトリックな方法における仮定の下に伏在している。

我々の研究室で実施した最近の慢性毒性試験において，臨床化学の統計的検定方法は，この章の最初に示されている決定樹に従って選択されている。このアプローチにより採択される方法は表11にまとめられてる。これはどんな方法が望ましいかについて指針として役立つ。更に詳細な論議はMartin等の著書（48）に見出される。

TABLE 10. *Association of changes in biochemical parameters with actions at particular target organs*

| Tests | Heart | Lung | Kidney | Liver | Bone | Intestine | Pancreas |
|---|---|---|---|---|---|---|---|
| Albumin | | | ↓ | ↓ | | | |
| Alkaline phosphatase | | | | ↑ | ↑ | ↑ | |
| Bilirubin (Total) | | ↑ | | ↑ | | | |
| BUN | | | ↑ | ↓ | | | |
| Calcium | | | ↑ | | | | |
| Cholinesterase | | | ↑ | ↓ | | | |
| Creatine phosphokinase | ↑ | | | | | | |
| Creatinine | | | ↑ | | | | |
| Glucose | | | | | | | ↑ |
| —GTP | | | | ↑ | | | |
| —HBDH | ↑ | | | ↑ | | | |
| LDH (L — P) | ↑ | ↑ | ↑ | ↑ | | | |
| Protein (Total) | | | ↓ | ↑ | | | |
| SGOT | ↑ | | ↑ | ↑ | | | ↑ |
| SGPT | | | | ↑ | | | |

Arrow indicates increase (↑) or decrease (↓) of chemistry values.

# 質疑応答

Q ： データの正規性が不確かとなり，ノンパラメトリック法の採用がすすめられるサンプルサイズの小さい場合というのはいくつぐらいか？

A ： ここでは１０以下の場合を意味していると思われる。

Q ： 表11でANOVA またはt-testとKruskal-Wallisの両方があるが，これは正規分布を仮定できるかどうかということで分けられているのか？

A ： 最初のチャート（決定樹）でもそういうことで分岐していますので，正規性が認められるかどうかということだと思う。
ただ，この表の（正規分布が前提とされている）検査項目でも正規性が保証されないのがいくつかあるようですし，連続性についても必ずしも保証されているといえないように感じる。

Q ： 生化学的パラメータは必ずしもめったに独立でなくていろいろ関連しているとしながら，個々の検査項目に統計的検定方法を指定していることに矛盾はないのか？

A ： 一つの個体について測定されたパラメータは当然互いに関連があるはずで，結論をだす時にその関連性は留意しておくべきことである。表に示されている統計的方法は，現状ではこういうやり方が採用されているというふうに理解している。従ってもっとbetterな方法があればその方がいいかも知れない。betterな方法として例えば，多変量解析的なやり方はどうかなと思っていますが。

Q ： 臨床化学は薬の安全性を考えた時，どういう位置を示しているのか？

A ： 現在，いくつかの検査項目については毒性試験ガイドラインで義務付けられている。臨床化学における検査項目は他の検査，例えば病理組織検査などと関連付けて（どの臓器に影響がある時はどのパラメータが動くのかということが予め知られているので）薬の作用場所の見当をつけたり，作用の程度を評価する助けとなる。

Q ： 体重変化とはGainと解釈していいと思うが，著者は体重そのものに比べ体重変化を用いる方が有利であるとしているが本当に有利なのか？

A ： どちらが有利なのか疑問に思う。特に群分けが正しくなされていれば，なおさらどちらとも言えないと思う。

Q ： 私は体重そのものの方がいいと思っているが，他にどちらがいいかで意見のある方は？

A ： どちらがいいとは言えませんが，体重そのものの場合，投与直後に増加量がかなり減るとその後一定の増加をしても体重そのものの変化（対照との差）は続く。そのような差を見つけるためには増加をみる方がいいのかも知れない。

A ： 若い頃に薬物の影響として体重差が出ると，その後回復期に入ってもなかなか対照群の値まで戻らない。この場合，実体重そのままで有意差検定をやるとずっと＊が付くが増体重でゆくとそうはならない。そういうことから確かに増体重の方が（申請に？）有利な面はあるにはある。それを科学的にどうみるかは別だが。一方，大動物（ウサギ，イヌ，サル）では体重のことを念頭において群構成そのものはままならないので，こういう場合は，ある程度増体重をみておかなければいけないと思う。

TABLE 11. *Tests used in analysis of clinical chemistry data*

| Clinical chemistry parameters | Statistical test |
|---|---|
| Calcium | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| Glucose | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| Blood-urea-nitrogen | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| Creatinine | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| Cholinesterase | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| Total bilirubin | Kruskal-Wallis nonparametric ANOVA |
| Total protein | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| Albumin | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| GTP | Kruskal-Wallis nonparametric ANOVA |
| HBDH | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| AP | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| CPK | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| LDH | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| SGOT | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| SGPT | ANOVA, Bartlett's and/or *F* test, *t*-Test |
| Hb | ANOVA, Bartlett's and/or *F* test, *t*-Test |

# 毒性学者のための統計学

シャニー・C・ガット，カロール・S・ウェイル

（訳）桑山　典之

帝国臓器製薬（株）

血液学

臨床化学で述べたことは、慢性毒性試験における血液学的測定にもあてはまる点が多い。試験を評価する実践的なアプローチとして、どの方法が最も適切かが確信されるまでは、決定樹に従って解析されなければならならない。ある場合には測定値と母集団の分布が種間ばかりでなく、ある種の系統間でも変動すること（また、その値は経時的に変動すること）に留意し、安易に扱ってはならない。

これらのパラメーターの多くは、相互に関連し測定方法に密接に依存している。赤血球（ＲＢＣ）と平均血球容積（ＭＣＶ）は、コールターカウンターのような機器により直接測定されるので、ふつうは、パラメトリック法に適している。しかしながら、ヘマトクリット値は実際にはＲＢＣとＭＣＶから計算されるであろう。もしそうならば、両者に依存している。もし、ヘマトクリット値がＲＢＣとＭＣＶから計算されるかわりに直接測定されるならば、パラメトリック法で解析してよい。

ヘモグロビンは直接測定されるし、また、独立した連続量である。しかしながら、ヘモグロビンは同時に多くの形や構造（オキシヘモグロビン，デオキシヘモグロビン，メトヘモグロビン，等）が実際は存在しており、分布も典型的な正規分布でなく、むしろ多峰分布である。それ故にWilcoxon や多重順位和検定のようなノンパラメトリックな手法が用いられる。

白血球（ＷＢＣ）と型別白血球算定について考えると他の問題が生じてくる。総白血球数は典型的なパラメトリック分析を行える正規母集団である。しかし、ふつう型別白血球算定は、１００個を１つの組として、１つまたはそれ以上の組を手で数えることにより行われる。そこで、好中球の相対的な割合の結果はパーセントもしくは総白血球数にパーセントを掛けることにより、絶対的な型別白血球の数として報告される。このようなデーター，特に好酸球の場合（分布が正規分布に近づいていない）は、ふつうノンパラメトリック法で解析されなければならない。

Stastics for Toxicologists
P311-313
（訳）　ライオン（株）
鉾城　豊

病理組織学的変化の出現率

近年，慢性（および亜慢性）毒性試験の動物から得られる多数の臓器に関する病理組織学的観察の重要性が増してきている。処置または暴露した動物で，統計学的に有意な増加率がみられた変化のみに関心を持つことは正しくない（例えば，変化が処置動物において1例のみまたは極く少数の発生でも，それが稀なタイプかもしれない場合に”警告を発する（raises a flag)”ことがある）。しかしながら，多くの場合，処置動物の変化が対照動物より有意に悪いかどうかを決定する唯一の方法が統計学的評価であることは事実である。また，癌かどうかということが唯一重要なことではなく，変化の程度がもっとも興味のあることである。

典型的には，対照群と処置群との間で，あるタイプの変化の発生率を分母の数を修正し，カイ二乗検定またはフィッシャーの直接確率検定を用いて比較している。分母の修正は，最初に腫瘍を発見した動物より前に死亡した動物を実験者が考慮から除外することにより生ずる（対照群と処置群のいずれにおいても）。

二つの大きな論争上の疑問がそのような比較に含まれる：

（a）片側または両側のいずれの分布を基にすべきか
（b）多重比較での影響と関係は何か

片側かまたは両側かの議論は，長期の発癌性試験の様な研究で仮説が正しく検定されているかどうかの疑問を生じさせる。

腫瘍発生率が対照群と処置群とで異なっているか？　この場合は二方向性の仮説であり，両側分布で検定している。

処置群の腫瘍発生率が対象群よりも高いか？　この場合は一方向性の仮説であり，片側分布のみを考えている。

この疑問の答えにはより理論的な関係がある。即ち，有意性は，片側の場合は両側より大きくなる（実際は正確に二倍）。例えば，フィッシャーの直接確率法により検定すると，両側で $p = 0.098$，片側で $p = 0.049$となる一連のデータは，それゆえ有意性

に異なった判断を下す。Feinstein(26) は非数学的な事項の背景に関して秀れた評論を提示している。即ち，片側または両側のどちらが正しいかは，試験の目的および考えられる結果を基に，事前の研究者の明確な定義に基づくべきであるというものである（もし二方向性の結果が可能なら，片側検定の使用は正当であろうか？）。

多重比較の問題はより難しいものである。慢性毒性試験では，雌雄および各動物種について多数の臓器の変化または腫瘍の発生率を検定している。それぞれの結果は，もし $p \leqq 0.05$ の信頼限界を越えたなら有意と判断している。ここで考えなければいけない事は " $p \leqq 0.05$ " の意味である。これは，第一種の過誤をおかす確率の水準である（帰無仮説を採らない時，我々の得る結果が不正確な結論となる場合）。それゆえ，試験から正しい帰無仮説を誤って捨ててしまう場合（false positive）の確率，５％の確率が存在するという事が認められている。実際の選択は第二種の過誤をおかす（安全ではない化合物を安全と見逃すこと）より低いものである（典型的には１％）。

これら二種類の過誤は互いに関連しあっている。即ち，第二種の過誤を小さくすると，第一種の過誤は大きくなる。この場合の問題は，もしそのような比較を多数実施するなら，にせの陽性結果を見出す機会を繰り返し持つことになる点にある。この場合，変化もしくは腫瘍の比較について，一つの試験でその有意性を70回以上も検定するかもしれず，にせの陽性を示す値は大きなものになる。

この誇張されたにせの陽性率の程度，そして，どの様にしてこの影響を減少させるかは，変化の大きさから評価されている。Salsburg(59)は国立癌研究所（ＮＣＩ）の発癌性試験における第一種の過誤の確率が20～50％であると評価している。一方，Fearsら (23,24)は 6～24％と評価している。

Salsburg(59)は見せかけのにせの陽性結果が，良い化学物質を追放してしまう事に関して懸念を表わしている（この事は産業界から広く支持されている）。Haseman(38)は規定された因子を用いる方法より更に賢明な判定法を試みた。しかしながら Salsburg は，化学物質を追放するための決定が単一の統計的な有意差に基づいている点について少なくとも二つの事を指摘している。

そのような結果の正しい使い方は何か？また逆に，どの様にして高くなりすぎた誤差率を制御できるか？

この多重比較の問題を処理するのに有効な統計手法がある。一つは，連続した多重比較を補正するためにBonferroni不等式用いることである(78)。この手法の欠点は真の陽性を正確に選択できない場合に，検出力の損失をいくらか伴なうことである。

二つめには，より優れた意志決定のために情報を使うことである。まず最初に，背景データの発生率を考慮すべきである［ B6C3F1 マウスとFischer 344 ラットの値は Fears らにより示されている（24）］。いくつかの背景データの発生率は高いので，結論を下す時にはこれらの臓器は意味の無いものである。

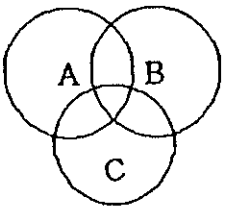
次に，臓器での単一な有意差をみるよりも，傾向をみるべきである。例えば，雌ラットの肝臓腫瘍で，(a) 対照群－ 3%,(b) 10 mg/kg 群－ 6%,(c) 50 mg/kg 群－17%, (d) 250 mg/kg 群－54%の値を示し，250 mg/kg 群の発生率のみがこの試験で統計的に有意であっても，各投与量間には用量相関性が示唆される。この様な傾向をみることは，結果を科学的に評価する本質的なステップであり，Tarone(67)が提示している様な傾向分析手法を役立てるべきである。

統計的に有意な発生率が単に偶然生じたかどうかを決定する他の手法は，変化量を同時に実施した二つ以上の対照群と比較することである。しばしば，一つの変量の平均が対照群のうちの一つとは異なるが，二つの対照群の平均値の間となる場合がある。もしそうなら，一つの対照群と統計学的有意差を比較することは，生物学的有意差を考え合わせる時に極めて疑問なこととなろう。

発癌性試験のデータを基にしたリスクアセスメントは近年広く認められている。高濃度の化学物質における実験動物の腫瘍発生率が，より低い実際濃度でのヒトにおける発癌リスクの予測に用いられている。発癌性試験のデータを予測または外挿する多数の手法があるが，これらの内，普遍的に認められたものは無く，評価のたよりなさの程度には数段階ある。リスクアセスメントに関する部分はひどく複雑であり，この話題を充分に探究し展開するには，ここで使えるよりかなり多くのスペースが必要であろう。GaylorとShapiro(32) のように，いくらかの概論は有効なものである。

## Bonferroniの不等式について（吉村先生の補足説明）

Bonferroni不等式の概念を面積を用いて表わすと以下の様になる（ここでは，Aの面積を面｛A｝と記号化することにする）。すなわち，３つの円が重なったものを考えた場合；




1．AとBとCの３つを合わせた面積として面｛A∪B∪C｝は，
面｛A∪B∪C｝≦面｛A｝＋面｛B｝＋面｛C｝となる。

2．重複した部分を除いた真の面積は，
≧（面｛A｝＋面｛B｝＋面｛C｝）
－（面｛A∩B｝＋面｛B∩C｝＋面｛A∩C｝）となる。

２式がBonferroniの不等式である。

これを確率として考えると，

$$Pr\{\max(TA,TB,TC) > a\} \geq (Pr\{TA>a\} + Pr\{TB>a\} + Pr\{TC>a\}) - (Pr\{TA>a,\ TB>a\} + Pr\{TB>a,\ TC>a\} + Pr\{TC>a,\ TA>a\})$$

となる。

# 毒性学者のための統計学

S.C. Gad and C.S. Weil(1982): Statistics for Toxicologists. In "Principles and Methods of Toxicology" edited by A.W. Hayes (pp 313−315).

訳: 飯田博司
日本ベーリンガーインゲルハイム(株)

## 生殖試験

化学物質の生殖に対する毒性効果を評価することは増々重要になってきた。他の関連性の強い試験(後に検討する催奇形性試験、優性致死試験など)に加えて生殖試験は今では慢性毒性試験を実施する際に並行して行なわれるものとなっている。

生殖に関する試験すべてについて留意すべき点のひとつに適切な標本単位はどういう性質のものかという問題がある。換言すると、このような試験での適切なNとは何か−−個々の仔動物数だろうか、それともlitter数だろうか。幸いにも今日では仔動物数をNとして用いる前者の考えかたは不適当であり、観察された差異の有意性を不当に大きくするだけのものだという見解が広く受け入れられている(73)。このような試験での真の効果は実際には化学物質を投与又は暴露した雌、あるいは投与雄と交配した雌に現われる。ある母動物とその胎仔の発生に起きる事柄は、他の母動物とその胎仔に起きる事柄とは生物学的に独立している。しかし1つのlitter内での個々の胎仔についてはそのようにいうことはできない。すなわち1つのlitter内では1個のメンバーの死や変化は他のメンバー全員に様々なかたちで影響する。いいかえればすべての仔動物にたいする影響は1匹の母動物から生まれる仔動物どうしについては似かよっているが、それは他の母動物から生まれる仔動物には異なった形で現われるか、あるいは現われないかもしれない。

Oser and Oser (55)の定義によると生殖試験には重要な4つの基本変数がある。第一に受胎率(Fertility index, FI)、これは交配に供した動物(雌を雄と同居させること)のうちで着床部位の存在を検査して、妊娠を確認した雌動物の百分率である。第二に妊娠率(Gestation index, GI)、これは膣栓の落下や膣スメア中の精子の鏡検により交尾を確認した妊娠動物のうち生存litters(少なくとも1匹の生仔のあるlitters)を有する妊娠動物の百分率である。この妊娠率と関連する2つの変数がある。すなわち一つはlitter当たりの産仔数であり、他は全産仔に対する死産仔のlitter当たりの百分率である。第三に生存率(Viable index,VI)、これは産仔のうち少なくとも生後4日間生存した仔動物のlitter当たりの百分率である。4つの基本変数の最後は、哺乳率(Lactation index, LI)、これ

は４日令で生存していた仔動物のうち離乳まで生存した仔動物のlitter当たりの百分率である。哺乳期間はラットとマウスでは以前から生後21日間としている。もう１つの変数としてlitter毎の平均仔体重増加量をこの試験の変数として取り入れるのが適している。

適切な例数についてはteratologyの節で述べるが、Ｎとして最小10を与えるならこれらの変数のそれぞれについてt検定またはANOVAを用いて有意差検定を行なうことができるだろう。もしＮが10より小さい場合には中心極限定理にしたがうことが期待できないので、Wilcoxonの順位和(２群比較)又はKruskal-Wallis nonparametric ANOVA(３群以上の比較)を用いる必要がある。

催奇形性試験

生殖に関する試験で第一にあげられるのは投与動物の仔に対する出生障害や奇形発現の可能性について検索する催奇形性試験である。このような試験でのデータ解析においては考慮すべき点がいくつかある。

第一に例数(すなわちlitter数)をいくらにすれば効果を評価する際に適切な信頼水準を得られるかという点である。この例数Ｎは次式によって算出できる(31):

$$N = \frac{(t_1 + t_2)^2 \, S^2}{d^2}$$

$t_1$＝有意水準に対応する自由度Ｎ－１のt値(t表より求める)

$t_2$＝検出力に対応する自由度Ｎ－１の片側t値

s＝標本標準偏差(通常過去のデータから求める)。計算式は

$$S = \sqrt{\frac{1}{n-1} \sum (v_i - \bar{v})^2}$$

v＝対象としている変数(着床数など)

d＝平均値の差の許容範囲(例えば着床数の場合では0.5)

例数Ｎの計算には自由度∞のt値を用いることによっても良い近似値が得られる。

生殖試験の節で述べたようにさらに基本的な考慮事項として、多数の動物を用いれば平均値の平均の分布は正規型に近づくということがある。これが中心極限定理の意味することの１つである。すなわち個々のデータは正規分布していなくても、その平均値は正規型に近づく。例数が10あるいはそれ以上なら結果を評価するのにt検定のようなパラメトリック検定を用いてもかまわないくらいに正規型へ近づく。例数が10未満ならノンパラメトリック検定(Wilcoxon順位和検定やKruskal-Wallis nonparametric ANOVAのような)の方が適当

である。その他にもいくつかの方法が示されているが(46,53)、広く受け入れられたり利用されている方法ではない。ノンパラメトリック法で広く用いられているもののひとつにWilcoxon-Mann-Whitney検定がある。この検定法では、各群のデータは最初に昇順に並べられ、次に全データを(対照群、投与群両方の値をいっしょにして)順位付けする。同位(tie)のものについては平均順位を与える。こののち各群毎に順位の和を求め、Uを次式によって算出する

$$U_t = n_c n_t + \frac{n_t(n_t+1)}{2} - R_t$$

$$U_c = n_c n_t + \frac{n_c(n_c+1)}{2} - R_c$$

ここにNc、Ntはそれぞれ対照群及び投与群の例数、Rc、Rtはそれぞれ対照群及び投与群の順位和である。

2群比較の有意水準を知るためには、Uc又はUtのうち小さい方の値を表(Siegel,文献64など)に示された臨界値と比べる。

以上の考察及び方法を心に置いて、ここで催奇形性試験で出会う実際の変数をみてみよう。それらの変数は容易に2つのグループに分類することができる――すなわち致死に関する測定値と催奇形効果に関する測定値とである(31)。致死に関する測定値には(a)1妊娠動物当たりの黄体数、(b)1妊娠動物当たりの着床数、(c)1妊娠動物当たりの生存胎仔数、(d)1妊娠動物当たりの着床前損失率、(e)1妊娠動物当たりの胚吸収率、(f)1妊娠動物当たりの胎仔死亡率がある。催奇形効果に関する測定値には(a)1litter当たりの異常胎仔発現率、(b)1群当たりの異常胎仔発現litter率、(c)胎仔体重増加量がある。

Mann-Whitney U testは計数データに対して用いるべきであるが、百分率データに対してどの検定法を用いればよいかという選択に際しては、生殖試験の節で述べたのと同じ基盤にたって決める必要がある。

訳終わり

〈吉村先生の解説、訳者メモによる〉

適切な例数nを求める計算式について。

いま、例として$\sigma^2$が既知の場合の1標本片側検定を考える。

対照群の値の分布　$N(\mu_c, \sigma^2)$ ：既知

処理群の値の分布　$N(\mu_t, \sigma^2)$

帰無仮説$H_0$:　$\mu_t = \mu_c$

対立仮説$H_1$:　$\mu_t > \mu_c$

検定は $\dfrac{\bar{x} - \mu_c}{\sigma/\sqrt{n}} > Z_1$ ならば有意差あり。ここで$Z_1$は有意水準での正規分

布のパーセント点，nは標本の大きさ、$\bar{x}$はある試験で得た標本平均。

$\mu_t - \mu_c = d$において検出力を$(1-\beta)$にするにはnをいくらにすればよいか。

$$Pr\left\{ \frac{\bar{x} - \mu_t + \mu_t - \mu_c}{\sigma/\sqrt{n}} > Z_1 \right\} = 1-\beta \quad \text{になるように } n \text{を決める.}$$

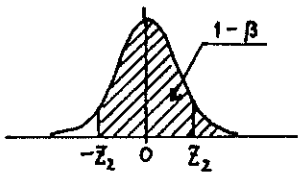
$$Pr\left\{ \frac{\bar{x} - \mu_t}{\sigma/\sqrt{n}} > Z_1 - \frac{d}{\sigma/\sqrt{n}} \right\} = 1-\beta$$

すなわち

$$Z_1 - \frac{d}{\sigma/\sqrt{n}} = -Z_2 \qquad (-Z_2 \text{は } 1-\beta \text{に対応する正規分布のパーセント点.})$$

$$Z_1 + Z_2 = \frac{d}{\sigma/\sqrt{n}}$$

$$n = \frac{(Z_1+Z_2)^2\sigma^2}{d^2}$$




ところで通常、$\sigma$は既知ではないので$\sigma$にかえてsを用いてt検定を行なう。$Z_1$, $Z_2$, $\sigma$をそれぞれ$t_1$, $t_2$, sに置き換えて

$$n = \frac{(t_1+t_2)^2 S^2}{d^2}$$

また2標本問題では次式が導かれる。

$$n = \frac{2(t_1+t_2)^2 S^2}{d^2}$$

## 〈質疑応答〉

(Q) パラメトリックとノンパラメトリックはどう使い分けるか。ここではN＞＝10ならパラメトリックと書かれてあるが、こういうことは普通に行なわれているのか？

(A) データ次第。litter sizeがある程度大きいことを想定しているのだろう。

(Q) データの分布型と検定について。

(A) もとのデータが正規分布でなくてもたくさんの平均値をとると正規分布に近づいてくるから、例数が等しければ正規分布で近似することはおかしくない。もとのデータ自身は正規分布していなくても、はずれ値(Outlier)がなく、例数がそろってい

て、平均値について吟味するなら分散分析や平均値の差の検定では有意水準は大きくは変わらない。正規分布からのずれが問題になるのはOutlierが存在するときである。Outlierがある場合パラメトリック手法を適用することは非常に危険である。

(A') 佐久間著「薬効評価1(4.3)」にt検定の頑健性の記述がある。理論分布を用いてのモンテカルロ法による吟味(Ratcliffe, 1968)によると、片側2.5%水準でのt検定では対称型なら3例以上、平坦型なら15例以上、凹型なら30例以上、非対称1山型なら50例以上、ポアソン分布のような極端な非対称型なら80例以上必要である。

[24回以降 基礎講座テキスト](3~4回の予定)

# 14 Methods and Concepts of Biometrics Applied to Teratology

D. W. GAYLOR

## I. INTRODUCTION

Some procedures are described which are used frequently in the statistical analysis of data from teratological studies. For the most part, the discussion here focuses on those species that produce litters with more than one offspring. There is a question of what teratological indices provide the most useful information. A discussion is presented on the proper choice of the experimental unit, the litter or fetus, for determining sample sizes. Owing to the large experimental variation often encountered in teratological studies, the desirability of replicated experiments is stressed. Measures of lethality and teratogenicity are discussed. Statistical tests for comparing treated and control animals are given. The use of fitting mathematical models to describe dose-response relationships is discussed.

## II. MEASURES OF LETHALITY

### A. Corpora Lutea, Implants, and Live Fetuses

If a compound is administered prior to nidation, the number of corpora lutea or number of implants may provide a measure of early lethality.

D. W. GAYLOR • National Center for Toxicological Research, Jefferson, Arkansas.

Frequently, compounds are not administered until after implantation, in which case only the number of live fetuses would serve as a measure of lethality.

Let $c_j$, $m_j$, and $f_j$ be the number of corpora lutea, implants, and live fetuses for the $j$th litter, where $j = 1, 2, \ldots, n$ is the number of pregnant mothers treated similarly. The best unbiased estimator of the average number of corpora lutea per litter is the simple arithmetic mean, $\bar{c}$, which is the sum of the corpora lutea for the $n$ litters divided by the number of pregnant mothers, $n$:

$$\bar{c} = \frac{1}{n}\sum_{j=1}^{n} c_j = \frac{1}{n}(c_1 + c_2 + \cdots + c_n)$$

The average number of implants and live fetuses per litter are computed in a similar manner.

The sample standard deviation for corpora lutea, $s_c$, is calculated by

$$s_c = \left[\frac{1}{n-1}\sum_{j=1}^{n}(c_j - \bar{c})^2\right]^{\frac{1}{2}}$$

Similar calculations apply for the sample standard deviation for implants and live fetuses.

Even though the number of corpora lutea, implants, or live fetuses are not normally (Gaussian) distributed, the Central Limit Theorem states that means of nonnormal variables rapidly approach normality (Mood, 1950). Thus, approximate confidence limits on the mean number of corpora lutea can be calculated when $n \geq 10$ by $\bar{c} \pm ts_c/\sqrt{n}$, where $t$ is obtained from standard statistical $t$-tables with $(n-1)$ degrees of freedom for the level of confidence selected. This interval includes the true average number of corpora lutea, at the chosen level of confidence, that would be expected from all litters treated under similar conditions. Approximate confidence intervals for the average number of implants and average number of live fetuses per litter are calculated in a similar manner. The approximation is reasonably good for $n \geq 10$, but this does not mean that 10 is necessarily an adequate number of litters. The number of litters required depends upon the precision desired and the variance.

For example, suppose that the numbers of live fetuses in a group of 16 litters are: 9, 12, 10, 8, 7, 12, 11; 13, 8, 7, 11, 12, 10, 10, 12, and 8. The average litter size is 10 and

$$s_f = \left[\frac{1}{15}\left[(9-10)^2 + (12-10)^2 + \cdots + (8-10)^2\right]\right]^{\frac{1}{2}} = 1.97$$

The 95% confidence interval for the mean number of live fetuses is $10.0 \pm 2.13(1.97)/\sqrt{16} = 10.0 \pm 1.05$.

The median sample size needed to achieve a given precision of $\pm d$ is given

by $n = t^2 s^2/d^2$, where $t$ has the desired level of confidence and $n-1$ degrees of freedom. The exact degrees of freedom, $n-1$, cannot be determined until the sample size, $n$, is determined. As a first approximation the $t$-value with infinite degrees of freedom can be used. For example, if it is desired to determine the average number of live fetuses within ±0.5 with 95% confidence, then $t = 1.96$. Using the estimate of the standard deviation for litter size of $s_f = 1.97$, the approximate median number of litters required is

$$n = \frac{(1.96)^2(1.97)^2}{(0.5)^2} = 59.6 \quad \text{or} \quad 60 \text{ litters}$$

This procedure provides an estimate of sample size which is adequate to achieve the desired precision about half of the time.

To be more certain that the sample size is adequate, the required number of litters is calculated by

$$n = \frac{(t_1 + t_2)^2 s^2}{d^2}$$

where $t_1$ is the $t$-value with $n - 1$ degrees of freedom corresponding to the desired level of confidence, and $t_2$ is a one-sided $t$-value with $n - 1$ degrees of freedom corresponding to the probability that the sample size will be adequate to achieve the desired precision. As a first approximation, $t$-values with infinite degrees of freedom can be used. In the previous example, if a probability of 90% is desired that the sample size is adequate to estimate the average litter size within ±0.5 with 95% confidence, then the estimated number of litters is

$$n = \frac{(1.96 + 1.28)^2(1.97)^2}{(0.5)^2} = 163 \text{ litters}$$

The discussion above on measures of lethality applies to the simple situation where a group of $n$ pregnant animals is treated similarly under a given set of conditions at some point in time. The total number of animals available for experimentation at that time may be randomly assigned to different groups (e.g., different dosage levels, different routes of administration of a chemical). The estimates of the average responses and confidence limits apply only to the existing set of laboratory conditions.

If an experiment is replicated (repeated) at $r$ different times or at $r$ different laboratories, a broader base for inferences is provided. If $\bar{f}_k$ represents the average for the mean number of fetuses per litter from the $k$th replicate ($k = 1, 2, \ldots, r$), then an overall estimate of the average number of fetuses per litter is given by

$$\bar{f} = \frac{1}{r} \sum_{k=1}^{r} \bar{f}_k$$

This estimate of $\bar{f}$ gives equal weight to each replicate. If the sample sizes, $n_k$, vary considerably among replicates and if the difference in the average responses among replicates is relatively small, it may be desirable to use an estimate of the average weighted by the number of litters per replicate. The calculation of the variance and confidence limits for these situations are considerably more complicated (Anderson and Crump, 1967).

The overall average number of corpora lutea or implants per litter from replicated experiments is treated in the same manner.

### B. Percentages: Preimplantation Loss, Resorptions, and Dead Fetuses

The reason for using percentages is an attempt to obtain more precise measures of lethality by adjusting for variability in the number of corpora lutea or implants from litter to litter.

If a compound is administered early in gestation, the percentage of preimplantation loss may provide a measure of early lethality. This can be estimated by subtracting the number of implants from the number of corpora lutea, expressed as a percentage of all corpora lutea, or as the percentage of corpora lutea which result in implants.

The percentage of implants that resorb is an appropriate measure of embryolethality when a compound is administered after implantation occurs. However, this quantity may not be appropriate if a compound is administered before implantation, as an increase in the percentage of resorptions may result from a reduction in the number of implants.

If a distinction is made between late fetal death and resorptions (embryo death), then the percentage of the implants resulting in fetal deaths can be calculated. Again, this may not be an appropriate measure of fetolethality if a compound is administered prior to implantation. The total lethality is represented by the percentage of implants resulting in either embryolethality identified by resorptions or late fetal death or equivalently by the percentage of live fetuses.

For all percentages in this section, the average percentage is simply the total of the percentages divided by the number of litters. For example, the average percentage of resorptions in a group of $n$ similarly treated mothers is

$$\bar{R} = \frac{1}{n}\sum_{j=1}^{n} R_j$$

where $R_j$ is the percentage of resorptions for the $j$th litter.

The sample standard deviation for the percentage of resorptions is

$$s_R = \left[\frac{1}{n-1}\sum_{j=1}^{n}(R_j - \bar{R})^2\right]^{\frac{1}{2}}$$

Where the number of litters is 10 or more, a reasonably good approximation for a confidence interval on the true percentage of resorptions is given by $\bar{R} \pm ts_R/\sqrt{n}$, where $t$ has the chosen level of significance and $n - 1$ degrees of freedom. Similarly, approximate confidence intervals can be calculated for the percentage of preimplantation loss or the percentage of dead fetuses. A situation that may arise, particularly for the percentage of dead fetuses, is that there may be no fetal deaths, giving $D_j = 0$ for all litters and $s_D = 0$. A confidence interval then could be based on the binomial distribution (see, e.g., Mood, 1950) for zero deaths out of $\Sigma m_j$ implants.

In an experiment replicated $r$ times, an estimate of the overall average percentage of resorptions per litter is given by

$$\bar{\bar{R}} = \frac{1}{r} \sum_{k=1}^{r} \bar{R}_k$$

where $\bar{R}_k$ represents the average for the $k$th replicate. Alternatively, an estimate of the average weighted by the number of litters per replicate may be used. The variances of these estimators are discussed by Anderson and Crump (1967).

The overall average for the average percent of live fetuses or dead fetuses per litter from replicated experiments can be computed in a similar manner.

### C. Percent of Litters with Resorptions or Dead Fetuses

In Section II.B the percentage of resorbed or dead fetuses was calculated for each litter and then averaged to obtain a measure of lethality. An additional measure is the percent of litters with resorbed or dead fetuses. Generally, if the percentage of resorptions per litter is high, it would be expected that the percentage of litters with resorptions would also be high. Thus, these two measures are usually highly correlated and reflect nearly the same thing—the extent of resorptions. However, it is possible that the average percentage of resorptions per litter could be quite high as a result of being concentrated in only a few litters. Thus, calculating the percent of litters with one or more resorptions provides a measure of whether resorptions are spread over litters or concentrated in a few litters; similarly, for dead fetuses.

Approximate confidence limits for the proportion of litters with resorbed or dead fetuses could be obtained from the binomial distribution. This is only a rough approximation, as a condition required for application of the binomial distribution is that each litter must have an equal probability of containing a resorbed or dead fetus. This condition undoubtedly is not satisfied, since litter sizes vary, which would result in differing probabilities of obtaining a resorbed or dead fetus.

## III. MEASURES OF TERATOGENIC EFFECTS

### A. Percent of Abnormal Fetuses

Let $a_j$ represent the number of fetuses in the jth litter possessing a certain type of anomaly. This anomaly may be a specific type, such as cleft palate. Or, $a_j$ may represent a group of anomalies, such as skeletal malformations. Or, $a_j$ may represent the number of fetuses in the $j$th litter with any anomaly. In calculating the percentage of abnormal fetuses per litter, it is necessary to divide $a_j$ by the number of fetuses, $n_j$, examined for that particular anomaly. For certain gross anomalies, all live fetuses are generally examined, so that $n_j = f_j$. However, for certain soft tissue anomalies, sectioning of organs may be required, or for skeletal anomalies special staining is required. It is common practice to select part of a litter for special staining for skeletal defects. Thus, the number of fetuses examined for a particular anomaly or group of anomalies is often less than the number of live fetuses in the litter. Thus, the percentage of animals with anomalies for the $j$th litter is calculated by

$$A_j = \frac{a_j}{n_j} \times 100\%$$

Unless an equal number of fetuses are examined for soft tissue anomalies and skeletal anomalies, the value of $n_j$ will not be the same for each type of defect. The average percentage of anomalies per litter is

$$\bar{A} = \frac{1}{n} \sum_{j=1}^{n} A_j$$

where at least one fetus from each of $n$ litters is examined for the anomaly. The sample standard deviation for the percentage of anomalies per litter is calculated by

$$s_A = \left[\frac{1}{n-1} \sum_{j=1}^{n} (A_j - \bar{A})^2\right]^{\frac{1}{2}}$$

For 10 or more litters, fairly good approximate confidence limits are given by $\bar{A} \pm ts_A/\sqrt{n}$, where $t$ has the desired confidence level and $n - 1$ degrees of freedom. A situation that may arise is that there may be no anomalies of a particular type, giving $A_j = 0$ for all litters and $s_A = 0$. A confidence interval then could be based on the binomial distribution (see, e.g., Mood, 1950) for zero anomalies out of $\Sigma n_j$ fetuses examined.

In general, observing the percentage of abnormal fetuses per litter and the percentage of resorbed and dead fetuses per litter should give a sufficient indication of the fetotoxicity of a compound. However, it is quite possible that

there is a correlation between the percentage of abnormal fetuses and resorbed or dead fetuses. For example, the abnormal fetuses may tend to have a higher death rate. Thus, the percentage of abnormal fetuses per litter may not appear to be unusual. Therefore, it also is advisable to include the percentage of normal life fetuses per litter in any statistical analyses.

For experiments replicated $r$ times, the overall average percentage of anomalies per litter is

$$\bar{A} = \frac{1}{r} \sum_{k=1}^{r} \bar{A}_k$$

where $\bar{A}_k$ is the average number for the $k$th replicate. Again, weighted averages could be considered for percentages of abnormal fetuses (Anderson and Crump, 1967).

### B. Percent of Litters with Abnormal Fetuses

In addition to calculating the average percentage of abnormal fetuses per litter, it is of interest to determine if anomalies occur throughout the litters or if they are concentrated in a few litters. This is accomplished simply by calculating the percentage of litters containing at least one abnormal fetus. This calculation can be based on a specific type of anomaly or can include all anomalies. Crude confidence limits could be obtained by applying the binomial distribution, but these limits are only rough approximations, as the probability of observing an anomaly in a litter changes, depending on the number of fetuses examined for that anomaly.

## IV. FETAL WEIGHT

Whereas reduction in fetal weight due to treatment by a compound may be difficult to interpret biologically, fetal weight appears to be a very sensitive measure of toxicity. This may be due in part to the fact that weight is a continuous variable, as opposed to a discontinuous variable such as observing only whether or not a fetus is dead or possesses a particular anomaly. There is more statistical information in a continuous variable such as fetal weight. In general, it is advisable to keep male and female weights separate. If $w_{ij}$ represents the fetal weight for the $i$th live male (or female) fetus in the $j$th litter, the best estimate of the average male fetal weight for the $j$th litter is

$$\bar{w}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} w_{ij}$$

where $n_j$ is the number of male fetuses in the $j$th litter. If a litter does not contain any live male fetuses, that litter is not included in the calculation. If $n$

litters contain live male fetuses, an estimate of the average male fetal weight for that group of litters is

$$\bar{w} = \frac{1}{n} \sum_{j=1}^{n} \bar{w}_j$$

If an experiment is replicated $r$ times, an estimate of the overall average male fetal weight is calculated by

$$\bar{\bar{w}} = \frac{1}{r} \sum_{k=1}^{r} \bar{w}_k$$

where $\bar{w}_k$ represents the average male fetal weight for the $k$th replicate. Again, weighted estimates could be considered (Anderson and Crump, 1967).

## V. COMPARISON OF CONTROL AND TREATED GROUPS

### A. Corpora Lutea, Implants, and Live Fetuses

A common objective of an experiment is to determine if a compound is lethal to embryos or fetuses or teratogenic when administered to laboratory animals under a particular set of conditions. A preliminary experiment may be needed to select a dosage regime which is not excessively lethal to the pregnant mothers. In a test comparing a treated group of animals with a control group of animals, the nonparametric Wilcoxon–Mann–Whitney $U$-test (see, e.g. Siegel, 1956) is recommended. This test makes no assumption as to the mathematical distribution of the response and still provides a fairly powerful test for determining if differences exist. When there are 10 or more litters per group, the ordinary $t$-test provides essentially the same results. A two-sided test is apropriate if one is interested in whether or not the compound causes either an increase or decrease in the response. In general, one would be interested in determining if a compound resulted in a decrease of corpora lutea, implants, or live fetuses; hence, a one-sided test would be utilized.

To illustrate the methods, consider the following results for the number of live fetuses obtained in 10 litters of mice treated with a compound compared with 10 control litters:

Control: 7, 8, 8, 10, 9, 5, 8, 7, 9, 9
Treated: 5, 8, 7, 7, 9, 8, 4, 7, 9, 6

The control group averages 8 live fetuses per litter and the treated group averages 7 live fetuses per litter. The question is: Does this 12.5% reduction in

litter size represent a statistically significant reduction in litter size due to the compound?

For the Wilcoxon-Mann-Whitney $U$-test, the data are ordered from lowest to highest values.

Control: 5, 7, 7, 8, 8, 8, 9, 9, 9, 10
Treated: 4, 5, 6, 7, 7, 7, 8, 8, 9, 9

The ordered rankings are

Control: 2.5, 7, 7, 12, 12, 12, 17, 17, 17, 20
Treated: 1, 2.5, 4, 7, 7, 7, 12, 12, 17, 17

where the average rank is assigned to tied values. The sum of the ranks for the control group is 123.5 and 86.5 for the treated group.

The value of $U$ is determined from

$$U_c = n_c n_t + \frac{n_c(n_c + 1)}{2} - R_c$$

or

$$U_c = n_c n_t + \frac{n_c(n_c + 1)}{2} - R_c$$

where $n_c$ and $R_c$ are the sample size and sum of the ranks for the control group and $n_t$ and $R_t$ for the treated group. In this example,

$$U_c = 10 \times 10 + \frac{10 \times 11}{2} - 123.5 = 31.5$$

or

$$U_t = 10 \times 10 + \frac{10 \times 11}{2} - 86.5 = 68.5$$

In determining the level of significance for a comparison of two groups, the smaller value of $U_c$ or $U_t$ is used:

$$U = \text{minimum of } U_c \text{ or } U_t$$

Critical values of $U$ are given by Siegel (1956) to obtain the level of significance. The value of 31.5 indicates a one-tailed significance level of approximately $P < 0.10$. That is, there is a 10% chance that the observed difference is not real but is due only to normal experimental variation.

For the parametric $t$-test of the same hypothesis, the variance among the control group is

$$s_c^2 = \frac{1}{9}\left[(7 - 8)^2 + (8 - 8)^2 + \cdots + (9 - 8)^2\right] = 2.00$$

and for the treated group is

$$s_t^2 = \frac{1}{9}\left[(5-7)^2 + (8-7)^2 + \cdots + (6-7)^2\right] = 2.67$$

An $F$-test with 9 and 9 degrees of freedom, $F = 2.67/2.00 = 1.33$, indicates that these two variances are not significantly different. Thus, a pooled variance with 18 degrees of freedom can be used for the $t$-test:

$$s^2 = \frac{18 + 24}{9 + 9} = 2.33$$

The $t$-test gives

$$t = \frac{8 - 7}{\sqrt{2.33(\frac{1}{10}+\frac{1}{10})}} = 1.46$$

which is significant at $P < 0.10$ for a one-sided test. Thus, the result may have arisen by chance alone with a probability of about 10%. If the variances of the treated and control groups are unequal, an approximate $t$-test can be employed (see, e.g., Anderson and Bancroft, 1952).

One might ask the question: How many litters would be required on the average in order to demonstrate if a difference of $d = 1$ fetus per litter is significant at the 95% confidence level? The formula for the number of litters, $n$, for each group is

$$n = \frac{2t^2s^2}{d^2}$$

where as a first approximation, $t$ is used with infinite degrees of freedom which would be $t = 1.64$ for a one-sided test. Thus

$$n = \frac{2(1.64)^2(2.33)}{(1)^2}$$

or 13 litters in each group.

In order to determine the number of litters, $n$, required to have a high probability, $P$, of being able to detect a difference of $d$, use

$$n = \frac{2(t_1 + t_2)^2s^2}{d^2}$$

where $t_1$ is the one-sided or two-sided $t$-value corresponding to the significance level for the test and $t_2$ is the one-sided $t$-value corresponding to a

significance level of $1 - P$. As a first approximation, $t$-values with infinite degrees of freedom (normal deviates) can be used.

A one-sided test is used where it only is of interest to establish if the treated group exhibits a change in one direction from the control group. A two-sided test is used where it is of interest to determine if there is a change in either direction, increase or decrease, from the controls. For example, the two-sided normal deviate corresponding to the 5% significance level is 1.96 and the one-sided normal deviate is 1.64.

In comparing data based on counts the square root of the counts are sometimes used to stabilize variances.

## B. Percentages: Preimplantation Loss, Resorptions, Dead and Abnormal Fetuses

For comparing percentages, some investigators have combined all the data from all litters and calculated the percentage of abnormal fetuses in the control groups and treated groups. The percentages of fetuses responding in the two groups have then been compared by the common $2 \times 2$ chi-square test. The procedure implies that each conceptus is an independent experimental unit with an equal probability of being defective. That is, it assumes that each conceptus in a litter or a conceptus in one litter is no more likely to receive more or less of the compound than a conceptus in another litter. Thus, pooling of data assumes no maternal effect. That is, it assumes that each pregnant mother handles a compound in a like manner, such that metabolism, excretion rates, and so on, are similar to the extent that the offspring can be thought of as coming from one large homogeneous litter. Obviously, such an assumption seldom would be warranted. Since the compound is administered to the pregnant mother, the mother or litter is the experimental unit. This topic has been addressed by a number of authors (Weil, 1970; Healy, 1972; Kalter, 1974; Staples and Haseman, 1974; Becker, 1974).

A recommended procedure is to compute $P_j$, $R_j$, $D_j$, $T_j$, and $A_j$; the percentages of preimplantation loss, resorptions, dead, resorbed and dead, and abnormal fetuses, respectively, for each litter. Percentages between control and treated animals then can be compared by the Mann–Whitney–Wilcoxon $U$-test (see, e.g., Siegel, 1956) or generally for sample sizes greater than 10 by the $t$-test (see, e.g., Anderson and Bancroft, 1952). A transformation commonly used for the $t$-test with proportions, $p$, is to use the arcsin $\sqrt{p}$, which nearly stabilizes the variances if the denominators of the proportions are of nearly equal size.

To illustrate the procedures for comparing percentages, consider the following data on resorptions in mice:

**Control Group**

| Litter ($j$) | Implants ($m_j$) | Resorptions ($r_j$) | Percent resorptions ($R_j$ |
|---|---|---|---|
| 1 | 7 | 0 | 0 |
| 2 | 9 | 1 | 11.1 |
| 3 | [illegible] | 2 | 22.2 |
| 4 | 8 | 1 | 12.5 |
| 5 | 6 | 0 | 0 |
| 6 | 10 | 0 | 0 |
| 7 | 7 | 0 | 0 |
| 8 | 9 | 1 | 11.1 |
| | | | 7.1 average |

**Treated Group**

| Litter ($j$) | Implants ($m_j$) | Resorptions ($r_j$) | Percent resorptions ($R_j$) |
|---|---|---|---|
| 1 | 8 | 1 | 12.5 |
| 2 | 8 | 0 | 0 |
| 3 | 9 | 1 | 11.1 |
| 4 | 6 | 2 | 33.3 |
| 5 | 8 | 3 | 37.5 |
| 6 | 9 | 0 | 0 |
| | | | 15.7 average |

For this example, $U = 16.5$, which is significant at approximately $P < 0.19$. There is not strong statistical evidence that the increased percentage of resorptions was due to the treatment, as there is approximately a probability of 0.19 that this difference is due to chance alone.

The problem above serves to illustrate the weakness of using too few litters in teratological studies. Even though the percentage of resorptions more than doubled in the treated group, this experiment was unable to provide strong statistical evidence that the effect was real.

The formula in the previous section could be used to determine approximately the number of litters required. In this case, using the transformation arcsin $\sqrt{p}$, where $p$ is the observed proportion, gives $s^2 = 0.0598$ for the example given above. If a one-sided $t$-test at the 5% significance level is used and it is desired to detect a doubling of resorptions from 7% in the controls to 14% in the treated group with a probability of at least 0.80, then the number of litters per group should be approximately

$$n = \frac{2(1.64 + 0.84)^2(0.0598)}{(0.384 - 0.268)^2} = 55 \text{ litters}$$

It is important to remember that statistical tests are only tools to be used by the scientist to interpret data. The quantities used to indicate toxicity may be correlated, requiring caution of interpretation. For example, the number of resorptions, dead, and abnormal fetuses may be correlated. The number of abnormal fetuses may decrease if there is a tendency for abnormal fetuses to be resorbed. In addition to the percentages calculated above, it is advisable to include the percentages of normal fetuses per litter in any analyses. Since the objective of most teratological studies is to look for adverse effects, there is a tendency to ignore measures of normalcy.

### C. Percent of Litters with Resorptions and Dead and Abnormal Fetuses

The techniques discussed in the previous section probably offer the best methods for detecting toxic effects. However, it is possible that the average percent of resorbed, dead, or abnormal fetuses may be quite high, simply because of a concentration of these conditions in only a few litters. For example, particularly small litters may have a high probability of being almost entirely resorbed, dead, or abnormal. Thus, a few litters could raise the average response considerably. Therefore, it is necessary to consider the distribution of a defect across litters. This can be accomplished by comparing controls and treated groups for the percentage of litters possessing a particular defect. A discussion of comparing two percentages is given by Snedecor and Cochran (1967). If the total number of litters is greater than 40, a chi-square test corrected for continuity of a 2 × 2 contingency table is suggested. The data may be displayed by a 2 × 2 table.

| | Number of litters | | |
|---|---|---|---|
| | No defects | One or more defects | Total |
| Controls | $a$ | $b$ | $n_c$ |
| Treated | $c$ | $d$ | $n_t$ |
| Total | $a + c$ | $b + d$ | $n$ |

The chi-square value is calculated by

$$\chi^2 = \frac{(|ad - bc| - n/2)^2 n}{n_c \times n_t \times (a + c) \times (b + d)}$$

This value can then be compared to the values tabulated in a chi-square table with one degree of freedom to determine the level of significance of the difference between $b/n_c$ and $d/n_t$. For a one-sided test, compare $\sqrt{\chi^2}$ with a standard normal deviate.

For example, suppose that 4 out of 20 (20%) control litters contained abnormal fetuses and 12 out of 30 (40%) of the treated litters contained one or more abnormal fetuses. Thus, $a = 16, b = 4, c = 18, d = 12, n_c = 20, n_t = 30$, and $n = 50$. Then,

$$\chi^2 = \frac{(|16 \times 12 - 4 \times 18| - 25)^2 \times 50}{20 \times 30 \times 34 \times 16} = 1.3825$$

giving $\sqrt{\chi^2} = 1.18$, which is significant at approximately $P < 0.12$ for a one-sided test.

Again, the question can be raised: How many litters would be required to have a high probability of detecting doubling of the percentage of litters containing abnormal fetuses from 20% of the control litters to 40% of the treated litters? The approximate number of litters required, $n$, in each group in order to have a high probability of detecting a specified difference is given by

$$n = \frac{(Z_1 + Z_2)^2}{2(\arcsin\sqrt{P_t} - \arcsin\sqrt{P_c})^2}$$

where $P_t$ and $P_c$ are the expected proportions in the treated and control groups, $Z_1$ is the normal deviate corresponding to the significance level of a one- or two-tailed test, and $Z_2$ is the one-sided normal deviate corresponding to $1 - P$, where $P$ is the minimal desired probability of detecting the difference between $P_t$ and $P_c$. In this example, if it is desired to detect a difference between 20% and 40% with a probability of 0.8 using a one-sided significance level of 0.05,

$$n = \frac{(1.64 + 0.84)^2}{2(\arcsin\sqrt{0.4} - \arcsin\sqrt{0.2})^2} = 63 \text{ litters per group}$$

A word of caution is necessary in comparing the proportion of litters exhibiting a certain defect. The chi-square test is only approximate because it assumes that each litter has the same probability of containing a defect. Since litter sizes vary, this condition is not satisfied. For a minimal condition for using this test, it is necessary that the average number of implants per litter is approximately equal for the control and treated groups. This should pose no problem for compounds that are administered after implantation takes place during gestation. However, if the number of implants is different between the control and treated litters, the test may be meaningless, as the larger litters may have a higher probability of containing a resorbed, dead, or abnormal fetus.

### D. Fetal Weight

The fetal weights for the two sexes should be analyzed separately. Fetal-weight data generate a three-level nested classification: control versus treated groups, litters within groups, and fetuses within litters. Since the number of male or female fetuses per litter will not be equal, the analysis of variance which provides an approximate $F$-test for the difference between groups is somewhat complicated and is not presented here (see, e.g., Anderson and Bancroft, 1952).

Fetal weight appears to be a sensitive indicator of toxicity. It generally is a more consistent measure than the numbers of implants, resorptions, etc., and percentages of dead fetuses, abnormal fetuses, etc. Even though the biological significance of a reduction in fetal weight may be difficult to assess, it generally is an indication of the presence of other toxic effects.

### E. Replicated Experiments

The examples given generally indicate the need for more litters than are typically used in teratological experiments. Thus, a moderate number of litters may be required. Laboratory facilities and resources may be inadequate to accommodate a large number of litters simultaneously. This situation is easily resolved by conducting the experiment in blocks (replicates) over a period of time. Animals in each replicate would be impregnated on the same date and must be assigned at random to the treatment and control groups. For example, a replicate may consist of 10 treated and 10 control litters started on the experiment on a given date. This process is repeated at later dates until an adequate number of litters are obtained.

The replicated experiment has an inherent advantage over a study conducted at one time. Considerable differences in responses have been noted between groups of animals treated at the same laboratory at different times even though laboratory conditions were supposedly similar. Thus, an experiment conducted at only one time may lead to different conclusions than the same experiment conducted at a different time. Reasons for these differences may be many and unknown and unfortunately can only be attributed to the rather vague but well-known existence of experimental variation. The replicated experiment provides an opportunity to provide a broader base for inferences from the data, as the average results are likely to be more representative and reproducible since the data are averaged over a wider set of conditions. One does not lose precision in such an experiment, because the treated and control groups are still compared within a replicate under similar conditions.

The statistical analyses of replicated experiments become more complex by extending the $U$-test to nonparametric analysis-of-variance techniques (see,

e.g., Siegel, 1956) or extending the *t*-test to parametric analysis-of-variance techniques (see, e.g., Anderson and Bancroft, 1952). Experiments involving more than one strain, route of administration, or days of treatment during gestation again require more complicated statistical analyses likely to include analysis-of-variance techniques for comparing several groups simultaneously. Snedecor and Cochran (1967) provide a discussion for combining results from a series of 2 × 2 tables for comparing percentages.

## VI. MULTIPLE-DOSE EXPERIMENTS: DOSE-RESPONSE

It is common to utilize several dosages in a biological experiment in order to study the intensity of a response as a function of dose. The establishment of a dose-response curve provides strong evidence of a cause-and-effect relationship between the administration of a compound and the observed biological response. A dose-response curve also may provide some indication as to dosage levels that may only produce negligible biological effects.

There are a multitude of types of dose-response regression curves which may be fit to data. No attempt will be made here to discuss the techniques available. Many statistical texts discuss curve fitting. In biology, particular attention has been given to fitting curves to quantal (percentage) data by probit analysis (Finney, 1971).

## REFERENCES

Anderson, R. L., and Bancroft, T. A., 1952 *Statistical Theory in Research*, McGraw-Hill, New York.

Anderson, R. L., and Crump, P. P., 1967, Comparisons of designs and estimation procedures for estimating parameters in a two-stage nested process, *Technometrics* 9:499-516.

Becker, B. A., 1974, The statistics of teratology, *Teratology* 9:261-262.

Finney, D. J., 1971, *Probit Analysis*, 3rd ed., Cambridge Univ. Press, London.

Healy, M. J. R., 1972, Animal litters as experimental units, *J. Roy. Stat. Soc., Appl. Stat.* C21:155-159.

Kalter, H., 1974, Choice of the number of sampling units in teratology, *Teratology* 9:257-258.

Mood, A. M., 1950, *Introduction to the Theory of Statistics*, McGraw-Hill, New York.

Siegel, S., 1956, *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York.

Snedecor, G. W., and Cochran, W. G., 1967, *Statistical Methods*, Iowa State Univ. Press, Ames, Iowa.

Staples, R. E., and Haseman, J. K., 1974, Selection of appropriate experimental units in teratology, *Teratology* 9:259-60.

Weil, C. S., 1970, Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis, and carcinogenesis, *Food Cosmet. Toxicol.* 8:177-182.

●取扱い代理店

株式会社 サイエンティスト社
東京都千代田区神田駿河台3丁目2番地
〒101 山崎ビル TEL 03-253-8992

NEC
NECパーソナルコンピュータ
PC-9800シリーズ

# 理・薬 医・生物 学万能

データ入力から
報告書作成まで
研究作業を強力に
バックアップ。

## PC-9800シリーズ用 MUSCOT統計解析 120,000円

### 販売代理のご案内

このたび、小社では㈱ワイ・デー・ケー（NEC特約店）およびユックムス㈱（医・薬学関係のソフトウエア開発の専門会社）と提携し、理・薬・医・生物学におけるデータの統計処理業務のために新開発されたソフトウエア「MUSCOT統計解析」（PC-9800シリーズ用）を販売する運びとなりましたので、ご案内申し上げます。

MUSCOT統計解析（MUltiple Statistics COmparing Tools）はNECのパソコンPC-9800／E／F／M2を用い、MS-DOS Ver2.0言語上で書かれた、数値標本やカテゴリー標本の統計解析を多面的に行なうソフトです。データの入力も合理的に操作でき、データに相関性がある場合は、自動的に多重比較法を選択するなど、データ入力から報告書作成まで一貫して効率的に行なうことができます。

この機会に是非、本ソフトをご利用頂きたくお願い申し上げます。また、適宜、デモンストレーションを企画しておりますので、ご希望の方は是非ご一報下さい。その都度、設定していく予定です。

＜特　徴＞

1. データ入力・訂正は合理的で簡単：スクリーンエディト機能により、画面を有効利用して確認作業も容易。ミスなく入力でき、コマンドも簡素化。
2. 自動的に手法を選択：多重解析作業を自動的に選択できるので、研究者の時間・手間が大幅に肩代りされ、研究作業をバックアップ。
3. フレキシブルなレポーティング：データの種類に合せて様々な作表が可能。タイトルや項目などをワープロ的に打ち込めて、簡単な編集機能付。

**パラメトリック〈検定〉**

1．平均、分散と1標本の検定／2．Studentのt検定／3．Aspin-Welchのt検定／4．Cochranのt検定／5．Bartlettの検定／6．一元配置分散分析(ANOVA)／7．Dunnettの多重比較検定／8．Duncanの多重比較検定／9．Scheffeの多重比較検定

**ノンパラメトリック〈検定〉**

1．Wilcoxon順位和検定／2．Mann-Whitneyの検定／3．Median検定／4．Van der Waerdenの検定／5．Kraskal-Wallisの検定／6．WilliamsのWilcoxon検定／7．Dunnett型の検定／8．Tukey型の検定／9．Scheffe型の検定／10．Jonckheereの検定／11．Sign検定／12．符号付Wilcoxon検定／13．Friedmanの検定／14．Page検定／15．Spearmanの順位相関係数／16．Kendallの順位相関係数／17．Kendallの一致性検定

**カテゴリー〈検定〉**

1．適合度の検定／2．$\chi^2$ 検定（2×2）／3．$\chi^2$ 検定（C×L）

**オートマティック統計手法**

1．StudentとCochranの2標本t-検定／2．Dunnettの多重比較検定／3．Duncanの多重比較検定／4．Scheffeの多重比較検定／5．Wilcoxonの順位和検定／6．WilliamsのWilcoxon型検定／7．Dunnett型の検定／8．Tukey型の検定／9．Scheffeの型の検定

### ●ソフトウエア●

MUSCOT統計解析では、10種類の機能を準備しています。

| 名　称 | 機　能 |
|---|---|
| 1．FILE | ファイル管理 |
| 2．EDIT | 統計データの入力・訂正 |
| 3．REDIT | 作表プログラム(REPOLC)の作成 |
| 4．REPOT | (REPOLC)実行 |
| 5．NCAL | 数値の計算と変換 |
| 6．SCAT | 標本分布を調べる |
| 7．PARA | パラメトリック検定 |
| 8．NONP | ノンパラメトリック検定 |
| 9．CATE | カテゴリー検定 |
| 10．AUTO MATIC | MUSCOT独自の機能で、統計解析の手法を自動的に選択し、簡単な操作で結果が得られる。 |

注）REPOLC：REPOrting Lunguage with Commands.

供給メディア：8インチ2D、5インチ2HD、5インチ2DDのうち1種類1枚をお届けします。


総発売元
YDK NEC特約店 株式会社 ワイ・デー・ケー

企画・開発
ユックムス株式会社

# 第23回定例会出席者名簿

日時：1985年7月13日(土)　14:00~17:00 定例会

場所：総評会館 2F会議室

11:00~14:00 基礎講座
(途中 昼食)

御出席頂いた先生方
・林　真（口立衛試）
・吉村 功（名古屋大）

1 小山　(東菱薬品工業)
2 永橋福太郎(クミアイ化学工業)
3 渡辺敏彦(科研製薬)
4 中嶋康彦(キッコーマン)
5 芳尾荘吉
6 中島信明(残留農薬研)
7 桑山典之(帝国臓器)
8 中嶋 圓(安評センター)
9 長谷文雄(グレラン製薬)
10 佐藤勝彦(ホーユー)
11 池田正己(日本ハイポックス)
12 高木 悟(ヘキストジャパン)
13 加藤正己(トーアエイヨー)
14 武政俊彦(ゼリア新薬工業)
15 山岡秀明(住友化学工業)
16 金子泰久(アップジョンファマシュティカルズ)
17 奥富庸雄
18 日浦謙一郎(ビーチャム薬品)
19 植村昌平(　〃　)
20 石塚修司(エスエス製薬)
21 松本 健(日水製薬)
22 大林繁夫(グレラン製薬)
23 朝野芳郎(エーザイ)
24 大川 豊(堀内伊太郎商店)
25 船山宣夫(　〃　)
26 阿部俊一(ミドリ十字)
27 重永敏明(大塚製薬)
28 安藤荘八(中外製薬)
29 滝沢 毅(日本ロシュ)
30 井野裕子(　〃　)
31 内田　(　〃　)
32 相馬義徳(　〃　)
33 中島敏秀(化研生薬)
34 安田栄一(佐藤製薬)
35 小林紀彦(バイエル薬品)
36 加納弘(台糖ファイザー)
37 三上正秋(東洋エタノールズ)
38 林　(　〃　)
39 御園等(三井製薬工業)
40 鷺坂豊(ライオン)
41 福本(参天製薬)
42 尾上正治(ヤクルト)
43 渡辺伸一(中外製薬)
44 堀江成光(参天製薬)
45 山下哲司(ロート製薬)
46 北山英太(日本新薬)
47 下井信夫
48 三浦昌己(東洋醸造)
49 今満裕(　〃　)
50 西原恵里(日本化薬)
51 玉島浚幸(大鵬薬品工業)
52 高橋みち子(日本臨床医学研)
53 縣直樹(三楽)
54 藤井祐一(津村順天堂)
55 小島 晴(養命酒製造)
56 茂田哲哉(日本レダリー)
57 佐藤七平(日本臨床医学研)
58 佐野正樹(生科技研)
59 横井義之(鐘紡)
60 福田武司(日生物化学センター)
61 松田和夫(ゼリア新薬工業)
62 大塚芳正(持田製薬)
63 奥村紘二(ヒューマンライフ)
64　〃　(　〃　)
65　〃　(　.　)
66 金谷木 稔(帝国臓器)
67 小杉典子(日本シェーリング)
68 永見俊之(日本農薬)
69 小山奇喜夫(鐘紡)
70 山口龍一(三菱油化薬品)
71 小山 薫(協和発酵)
72 北川行夫(サンド薬品研)
73 渡辺 実(　〃　)
74 河上喜之(東中研)
75 高橋昌三(日本チバガイギー)
76 小笠原定則(関東医科製薬)
77 米島　(帝人)
78 池田　(　〃　)
79 浅沼喜信(興和)
80 田中 健(日科技研)
81 菅井泉一郎(クミアイ化学工業)
82 大畑雅子(薬業時報社)
83 今井節夫(動物繁殖研)

[事務局だより]

一つの研究会をあずかり、運営していくと、人数が多くなり（現在、会員数は200人ちょっと）組織が大きくなる、ということだけでなく、活動の枠が広がるものだな、と思う。当会も、当初午後だけの定例会だったのが基礎講座を設けて、全日の例会になり、そのうち単行本の企画が出て、吉村先生、世話人の方々をわずらわせて、荒原稿から秘書へと作業が進んでいる。書名は、

『薬効・毒性データの統計解析
——事例研究によるアプローチ——』

と決定し、目次も、ほぼ決まってきた。一方、当会をきっかけにして、統計解析のソフトである「MUSCOT統計解析」の販売代理など。また、会報も、内容を充実していかねばならない。どんどん広がるような感じである。でも、枠の拡大に、それなりに対応しているから不思議である。どうやら、小生の能力も、少しずつ拡大しているらしい。

次回　第24回定例会（1985年10月13日(土)）の場所は、"グリーンホテル"です。

・基礎講座：Methods and Concepts of Biometrics Applied to Teratology
・定例会：（予定）
・データの正常範囲をどう考えるか
・GLP下での生データとは
・QAUからみた生データ
（初めて）・確率変数とは？

医薬安全性研究会　会報No.17
昭和60年7月31日発行
編集・発行　(株)サイエンティスト社
〒101東京都千代田区神田駿河台3-2
山崎ビル ☎03(253)8992 振替東京8-71335
印刷・製本　ナガノ印刷