日本統計学会誌
第22巻，第3号（増刊号），1993年
475頁～492頁

臨床試験の統計学的側面

広　津　千　尋*

Statistical Aspects of Clinical Trials

Chihiro Hirotsu*

While the general principles of data analysis apply also to clinical trials, such trials have yet special features. First there are many sorts of variation factors such as institutes, severity of illness and cause germs, which are fixed and indicative factor within a trial but act as if they were noise in the actual clinical treatments. We describe for the model based approach to this problem the generalized linear models including the proportional hazards model, logistic regression and regression models for ordinal categorical data. It requires, however, a careful examination of the model before extending the result beyond the trial and in some cases the double-blind randomized design based approach is recommended, which is described in § 2. Next in the analysis of clinical trials mostly concerned are the nonnormal distributions and the method based on the rank data can be applied for the ordinal categorical data, and vice versa. The analysis of repeated measures is a developing field of which we introduce generalized multivariate analysis of variance approach, two-stage mixed effects model and nonlinear mixed effects model. Other important issues discussed are various kinds of multiplicity problems and proving equivalence of a new drug with the standard.

臨床試験には一般的な統計解析の原理はすべてあてはめられる一方，いくつか特徴的なこともある．第一に施設，重症度，起因菌のように，一つの臨床試験内では母数的な標示因子と考えられるのに，実際の臨床処置においては必ずしも特定できず，誤差と考えざるを得ない，いわゆる変動因子が多数存在する．これに対処するモデルベーストの接近法としてここでは比例ハザードモデル，ロジスティック回帰，順序分類データに対する回帰モデル等の一般線形モデルについて述べる．ただし，解析の結果を当該臨床試験を超えて適用するにはモデルの十分な吟味が必要であり，二重盲検無作為化試験に基づくデザインベーストの接近法が必要であることも多い．それについては第2節で述べる．次に臨床試験では非正規分布を取り扱うことが多いが，分布形を仮定しない順位データに基づく方法は順序分類データにも用いることができその逆もいえる．経時測定データの解析は現在発展中の分野であるが，これに関しては一般化多変量分散分析模型に加え，2段階混合模型，非線形混合模型を紹介する．その他の重要な問題として推論の多重性の問題，および新薬と標準薬の同等性を証明するための統計的方法について述べる．

1. はじめに

臨床試験には一般的な統計的推測の原理はすべてあてはめられる一方，臨床試験に特徴的な問題も多く，独自の展開を産んでいる．

Cox (1982)でも述べられているように，処理比較の構造はむしろ簡単で多くの場合2もしくは3処理の比較が対象となるが，人（被験者）が繰り返しをなすため，年齢・性別などの人口

論文受付：1993年3月　改訂受付：1993年6月　受理：1993年6月
* 東京大学工学部，〒113　東京都文京区本郷 7-3-1

統計的要因の他，過去の病歴，重症度など考慮すべき多くの共変量が存在する．さらに施設（および担当医），あるいは起因菌のように事後的には同定できその影響を評価できても，実際の臨床現場では事前に同定できずあたかもノイズのように働く数多くの要因が存在する．このように一つの臨床試験内では母数的な標示因子と見なせるが，その結果を一般に拡大する際には制御できない変量因子と考えざるを得ない因子は品質管理の分野でも誤差因子（noise factor）あるいは変動因子（variation factor）と呼ばれ，その取扱いが最近一つの重要なトピックとなっている．

被験者が時間的に試験に組み入れられ，長期生存時間研究での計画的な右側打ち切り（right censoring）の他に，転院などによる脱落，副作用による治療変更，医師の処置違反，患者の治療不順守，あるいは事後的に判明する不適格など不完全例が多く発生するのも臨床試験の一つの大きな特徴である．これらにより生じ得る処理比較の偏りを極力おさえるため，試験計画および解析で様々な工夫がなされる．

さらに困難な問題はそもそも解析特性値として何を採用すべきかが必ずしも明確でないことである．比較的効果の発序が明確と思われている感染症の領域ですら，菌の消長と臨床症状の改善は必ずしも相関しないし，単独菌感染と複合菌感染の違い，あるいは急性と慢性を区別すべきか否かなど議論が絶えない．高血圧症では血圧，また高脂血症では体内総コレステロール量，HDL（善玉コレステロール），LDL（悪玉コレステロール），そしてトリグリセライドなどの計量特性値が測定され解析されるが，これらも脳血管障害や心臓病など本来の疾患に対する代用特性（surrogate endpoint）に過ぎない．さらにこれらのように本来の疾患と相関が強く，再現性，安定性もあって予測力の高い代用特性の得られない領域も数多くある．たとえばうつ病などの精神性疾患ではこのような信頼すべき計量特性値は得られず，様々な心因的スコア（の総合）が解析されるが，その精度および本来の疾患の改善度との関連はまた議論を要するところである．

良く効く薬には多かれ少なかれ何らかの副作用が伴うのが普通であるが，改善度と副作用の程度をどうバランスさせるかも重要な問題の一つである．我が国では改善度と副作用のデータをそれぞれ集計し解析する他，両者を勘案した主治医の判定である有用性が重要な情報として解析されている．有用性についてはその定義が不明確で担当医の主観が入り得るとの批判もあるが，有意な改善と副作用が同一の患者に出現したのか否かは臨床的に重要な意味があるのに別個の集計ではそれが失われてしまうこと，また，たとえば感染症における抗生剤のように救命治療を目的とする場合と，高脂血症剤のように特性値のコントロールを目的とする場合では同程度の副作用に対して重み付けが異なることを考えると，一人の被験者内で改善と副作用をバランス評価した有用性も重要な情報を提供すると思われる．そこで疾患ごとに，信頼すべき代用特性がありそれを解析対象とするのか，あるいはできるだけ基準を統一し，客観性を確保した上で担当医による有用性判定を解析するのかを定める必要がある．

生物学的データは計量特性値であっても一般に正規分布よりは裾が重く，対数正規分布やワイブル分布のように右に裾をひくことが多い．そこで非正規分布のための統計的方法が重要になる．このためノンパラメトリックな方法が使われることも多い．さらに，改善度や重症度のような順序分類データとして得られることも多く，それに基づく処理比較の方法やモデリングの方法がいろいろ提案されている．順序分類データはタイ（同順位）の多くある順位データとみなすこともできるので，順位に基づくノンパラメトリック法は順序分類データにも適用できるし，その逆もいえる．

多群比較の場合は一様性検定に替えて，目的に応じた多重比較法がよく用いられる．たとえば，治験薬群と対照薬とを比較する Dunnett 法，用量を上げたときに零水準に比べ有意な効果

の発序する水準を検出するためのWilliams法，同じく用量を上げたときにその前後で効果の差が大きく異なる点を検出するためのmax t法などがあり，効率的な有意確率計算のアルゴリズムの工夫や検出力の比較などが行われている．なお，高血圧症に関しては効果を見ながら用量を上下する独特な用量増減法が用いられるが，このように効果依存のストッピングルールを用いた場合の解析法はいまだよく解明されていない．

以上臨床試験の特徴を述べたがその多くは単に数学的な処理に留まらず，実質科学的に意義のある方策を探求するのに臨床医と統計学者の密接な協議が必要であるばかりでなく，被験者の協力や倫理的側面の考慮まで要請される．臨床試験はこのように極めて学際的な分野であるが，本論文ではとくに統計学関連事項に限って述べることとしたい．

2. 臨床試験の計画

動物による毒性試験やインヴィトロの薬理試験などのいわゆる前臨床を経た後，人での臨床試験はI～IV相の四つのステップを踏んで行われる．I相は少数の健常者（10名程度のボランティア）による安全性試験であり，II相は用量設定試験，III相が適切な対照薬をおいた比較試験，そしてIV相は市販後の調査ならびに再審査である．特に，稀な副作用は限られた例数による臨床試験では捉えられない可能性が強いので，IV相に負うところが大きい．II相は前期の比較的少数例による瀬踏み試験と，後期のより多数例による試験とに分かれる．以下ではこれらのうち特に，後期II相とIII相の比較臨床試験における統計的諸問題について述べる．

人による臨床試験ではセレクションバイアスと心因的プラセボ効果によるバイアスを避けるために，原則として無作為化二重盲検試験（randomized doubly blinded test）が行われる．二重盲検の名は被験者，担当医の両者が用量，あるいは治験薬，対照薬の別を知り得ないように無作為投与を行うことに由来している．

無作為化のレベルはいろいろである．とくに時間的均一性を意図したものとして置換ブロック法，バイアストコイントス法，最小化法などがある（Pocock, 1983）．置換ブロック法とは，被験者を組み入れ順に大きさ4ないし8のブロックに区切り，最初のブロックに比較すべき同数の処理をランダムに割り付け，その後のブロックではそれを順次巡回置換した割り付けを用いるものである．バイアストコイントス法は2処理A，Bの比較において第r番目の被験者の割り付け確率変数 Δ_r を

$$\Delta_r=\begin{cases}1, & r\text{番目の被験者が処理Aを受けたとき},\\ -1, & \quad〃\quad\ \text{B}\quad〃\end{cases}$$

と定義し，

$$S_r=\Delta_1+\cdots+\Delta_r$$

とするとき，与えられた $S_r=s$ に対し

$$\mathrm{p_r}=\{\Delta_{r+1}=1\}=\begin{cases}p, & s<0\\ 1/2, & s=0\\ 1-p, & s>0\end{cases}$$

であるような無作為化を行うものである（Efron, 1971）．ただし，$1/2<p<1$ であり，たとえばEfronは $p=2/3$ を薦めている．これにより常に割り付けの均等化がはかられるが，解析上この計画が被験者の反応（改善度）に依存していない（noninformative）ことが重要である．Cox（1982）はこの種の計画で得られた試験の解析について考察し，とくに

$$E(\Delta_{r+1}|S_r=s)=-\varepsilon s,\ 0.01<\varepsilon<0.1$$

である計画の具体的な解析法を述べている．最小化法は若干の重要な予後因子を，比較する処理間で強制的にバランスさせようとするものである．すなわち，注目する予後因子のアンバランスを測る損失関数を用意し，常に損失関数を最小にする割り付けを採用する．重要予後因子が既知である場合に，この方法によりその偏りを避けることができるが，確率化の要素が入らないため，置換ブロック法における並べ替え検定のような適切な統計処理の方法が得られないことが難点である．なお，一般に臨床試験は複数の施設にまたがって行われるため，ここに述べた無作為化法を行うには割り付けセンターによる電話登録法などを用いることが必要になる．

さて，我が国で実際によく行われているのは施設をブロックとする局所無作為化である．これは施設ごとに大きさ4～8のブロックを設けその中で処理を無作為に割り付けるものである．各施設は並行して試験を進めるので，これにより時間的な無作為化も図られることになる．ところで試験に組み入れる施設数および施設当りの被験者数については多くの議論があり，日米の臨床試験遂行上での一つの大きな差異であることが指摘されている．すなわち，日本では50～100あるいはそれ以上の施設を用い，一施設で少ない場合には例数1，2ということが多いのに対し，米国では優れた少数の施設を用い，一施設当りに施設間差を推定できる程度の例数を確保することが行われている．これは変動因子である施設を，日本では臨床の現場に則して変量因子とみなし，米国は標示因子（母数）と考えていることに他ならない．背景として日本は全土で比較的均一性が期待されるのに，米国では地域差が無視し得ないくらいに大きいことが挙げらているが，それを裏づける定量的解析は未だ不十分である．この両者はどちらが正しいというものでもなく，それぞれに長所，短所がある．日本流では施設間差は処理比較に当ってノイズとして作用するため比較精度が減ぜられる．米国流では施設間差を推定すると同時にその変動を除去した精度のよい処理比較が可能であるが，治験対象の母集団が実際の治療を受ける母集団と少なからず食い違うrepresentativenessの問題がある．米国では結果の普遍性を確保するために複数の独立な試験を行うとされているが，やはり施設がよりselectiveである問題点は残ると思われる．日本でも施設ごとの局所無作為化に対応して，局所的並べ替え検定が用いられることもあり，それによれば施設間差は除去される．しかしながらそれでは一例のみの施設は解析に組み入れられないなどの難点から実際上は完全無作為化を想定した解析がなされることが多い．新しい統計ガイドライン（厚生省，1992）で同等性の積極的証明（後述）が求められるようになったいま，比較精度とrepresentativenessについてより精密な議論が必要である．

以上，一般的な比較的短期の臨床試験の計画について述べたが，抗がん剤の生存時間研究のように長期（5～10年）にわたるものでは，患者の追跡や，効果あるいは副作用が判明した場合にそれをできるだけ速く実際の臨床に反映させるための中間解析が必要となり，そのための特別な機構が必要である．これについてはBuyse et al.（1988）を，また，生存・死亡や発症予防を目的とした大規模臨床試験についてはMeinert（1986）などを参照されたい．最近の日本のガン臨床試験で十分に吟味されたプロトコールに基づく場合でも，中間解析の結果副作用によりプロトコールの変更に至ったものが10～15％に達する他，明確な効果の差による中止も一例報告されている．なお，抗がん剤は第III相試験で腫瘍の縮小が確認された段階で認可し，その後第IV相で長期の生存時間比較研究を行う．これは他の臨床分野と大きく異なるところである．

薬剤の投与終了後に症状が元に復すような特別な場合には，適切なwash outの期間を設け，

同一被験者に2剤A，Bを適用することにより患者内比較が可能である．この場合，被験者をランダムに2分し，第1群にA→B，第2群にB→Aと適用するクロスオーバー法により薬効と時期効果を分離して推定できる．

3. ノンパラメトリックなアプローチ

臨床検査データは対数正規分布などの非正規分布に従うことが多く，処理比較には順位に基づくノンパラメトリック検定がよく用いられる．ここですべてのノンパラメトリック検定はバリディティロバストであるが必ずしもエフィシェンシーロバストではないことに注意する必要がある．たとえば2標本並べ替え検定は漸近的に t 検定と同等であり，正規分布もしくはそれに近い分布では効率が高いが，より裾の重い分布あるいは歪んだ分布に対しては急速に効率を落とす．正規分布の位置比較に対する局所最強力順位検定として Fisher-Yates スコア検定および Van der Waerden スコア検定がよく知られているがこれらの特性も同様である．ロジスティック分布の位置変化に対して最適な Wilcoxon スコア検定(順位和検定)は，これらに比べるとやや頑健性が高いが極端に裾の重い分布あるいは歪んだ分布に対してはやはり効率が低下する．もちろん他にもいろいろなスコア選択の可能性があり，歪んだ分布に対するものとしては極値分布の位置変化に対する Savage スコア

$$w_i=\frac{1}{n}+\frac{1}{n-1}+\cdots+\frac{1}{n-i+1}-1 \tag{3.1}$$

が知られている．Savage スコアは式 (3.1) が近似的に

$$w_i=\mathrm{loy}_e n-\mathrm{loy}_e(n-i+1)-1$$

と表されることからログランクスコアとも呼ばれ，むしろ指数分布の尺度母数に対する最適スコアとしてよく知られている．しかしながらこのようにいろいろなスコアの提案があるということは裏返していえば一つひとつは必ずしもエフィシェンシーロバストではないということであり，それが自由度1の線形スコア検定の限界となる．一方，一組のデータセットに繰返し異種のスコア検定を適用すれば第一種の過誤の増大は明白であり，少なくとも後期II相やIII相のような検証目的の試験では避けなければならない．そこで事前に適切なスコアを選択することが要求されるが，それはそもそもノンパラメトリック手法が必要とされる状況では必ずしも容易ではない．分布形を推定しつつ漸近的に最適スコアを選択する適応的な方法もあるが，それがよく機能するには可成りの例数を要する．そこで多自由度の手法として提案されているのが累積 χ^2 法および max χ^2 法である．その説明のために表1の例を考える．表2はそれを順位デ

表1　ラットにおける抗生剤投与後の半減期

投与量 (mg/kg/day)	データ (h)					平均
1.　25	1.55	1.63	1.49	1.53	2.14	1.688
2.　200	1.78	1.93	1.80	2.07	1.70	1.856

表2　表1の順位データ（小さいほうから順位1，2…とする）

	順位										
投与量 (mg/kg/day)	1	2	3	4	5	6	7	8	9	10	計（サンプルサイズ）
1.　25	1	1	1	1	0	0	0	0	0	1	$5=n_1$
2.　200	0	0	0	0	1	1	1	1	1	0	$5=n_2$
計	1	1	1	1	1	1	1	1	1	1	$10=n$

ータに表したものである．

表2において第 i 群からのサンプルが第 j 位となる確率を $p_{ij}(i=1,2\,;j=1,\cdots,10)$ と表すと，200 mg 群が 25 mg 群より統計的に大きいという一つの表現は

$$\frac{p_{21}}{p_{11}}\leq\frac{p_{22}}{p_{12}}\leq\cdots\cdots\leq\frac{p_{2n}}{p_{1n}} \tag{3.2}$$

で与えられる．ただし，式 (3.2) において少なくとも一つの不等号は厳密とする．たとえば背景分布が単調尤度比を持てば式 (3.2) が満たされる．ここで，ノンパラメトリック検定がエフィシェンシーの意味でもロバストであるためには式 (3.2) を満たす広い対立仮説の範囲で高い検出力を保持しなければならない．

いま，表2で 25 mg 群の 0，1 データを z_j とおく，

$$z_j=\begin{cases}1\text{，第 }j\text{ 位が 25 mg 群のとき，}\\0\text{，第 }j\text{ 位が 200 mg 群のとき．}\end{cases}$$

このときスコア w_j の線形スコア統計量は $\sum w_jz_j$ と表される．ここで単一のスコアシステム $\{w_j\}$ を考える代わりに，同時に九通り（一般には $n-1$ 通り）のスコアシステム，

$$\{0,1,\cdots1\},\quad\{0,0,1,\cdots,1\},\quad\cdots,\quad\{0,\cdots,0,1\}$$

を考え，そのそれぞれに対する線形スコア統計量の規準化二乗を $\chi_1^2,\chi_2^2,\cdots,\chi_9^2$ とおく．χ_J^2 は表2において，1，…，J 列および $J+1$，…，10 列をそれぞれプールして得られる 2×2 表に対する Pearson の適合度 χ^2 に他ならない．そこでこれを

$$\chi_J^2=\chi^2(1,\cdots,J\,;J+1,\cdots,10)$$

のように表わす．このとき累積 $\chi^2(\chi^{*2})$ および $\max\chi^2$ は次のように定義される．

$$\chi^{*2}=\sum\chi_J^2$$
$$\max\chi^2=\max\chi_J^2$$

χ^{*2} は次の展開式（一般の n で表わす）

$$\chi^{*2}=\frac{n}{1\times2}\chi_{(1)}^2+\frac{n}{2\times3}\chi_{(2)}^2+\cdots+\frac{n}{(n-1)\times n}\chi_{n-1)}^2 \tag{3.3}$$

で特徴付けられる (Hirotsu, 1986)．ただし，$\chi_{(J)}^2$ は Chebyshev の選点直交多項式の第 J 次多項式の各点をスコアとするスコア統計量の規準化二乗である．式(3.3)と Anderson & Darling (1952) の適合度統計量の直交多項式による漸近展開との類似性に注意して欲しい．式 (3.3) において最初の数項が支配的であることは容易に分かるが，$\chi_{(1)}^2$ は Wilcoxon 統計量の規準化二乗に他ならず，$\chi_{(2)}^2$ は Siegel-Tukey スコアのように分散の違いを検出するための統計量である．従って χ^{*2} は主として位置変化を検出し，それに分散の変化が重量されたような場合にも高い検出力を保持する．一方 $\max\chi^2$ はその構成からいって凸錐 (3.2) のエッジで高い検出力を持つので，対数正規分布や指数分布のように歪んだ分布に特に適している．

χ^{*2} は帰無仮説の下で1次，2次モーメントを合わせた χ^2 統計量の定数倍 $d\chi_\nu^2$ でよく近似されるが，必要に応じ Laguerre の直交多項式の展開により3次モーメントまで調整することもできる (Hirotsu, 1979, 広津, 1992a)．n を無限大にしたとき，ν は 3.45 に収束するがこれは線形スコア統計量の自由度1および適合度 χ^2 の自由度∞に比べ，χ^{*2} の指向性，多自由度性を

よく表している．一方，max χ^2 についてはその成分 $\chi_1^2, \chi_2^2, \cdots, \chi_{n-1}^2$ がマルコフ性を持つことから，正確な有意確率を計算するための簡潔なアルゴリズムが開発されている（Hirotsu et al., 1992）．max χ^2 は次々と 2×2 分割表で数え上げを行うこの方法により，片側検定にも使用される．

これら種々のノンパラメトリック検定のシュミレーションによる検出力比較は尾上他(1992)が行っている．その結果から，背景分布が正規分布もしくはそれに近い分布のとき t もしくは数値をそのまま用いる並べ替え検定，それよりやや裾が重く対称な分布に Wilcoxon 検定，Cauchy に近い分布には Kolmogorov-Smirnov 検定，歪んだ分布には max χ^2 検定が薦められる．そしてこれらすべての場合を通して χ^{*2} は安定した検出力を示すので，分布形に関する情報の乏しい時に χ^{*2} の使用が薦められる．

4. 順序分類データの解析

表 3 はある後期II相の用量設定試験の過程で得られたデータであるが，臨床試験においてはこのような順序分類形式のデータがよく採られる．

表 3　投与量別有用度

投与量 (mg/kg/day)	有用度 1.好ましくない	2.やや好ましくない	3.有用でない	4.やや有用でない	5.有用	6.極めて有用	計
1. AF 3	7	4	33	21	10	1	76
2. AF 6	5	6	21	16	23	6	77
計	12	10	54	37	33	7	153

このような表は表 2 において同順位が多数現れたものと見ることができ，解析の方針は前章とそれ程異ならない．よく用いられるのは平均順位を用いた Wilcoxon 検定，累積 χ^2 そして max χ^2 である．この場合の統計量および有意確率の計算はたとえば広津（1992b）を参照されたい．表 3 に適用した結果は表 4 のようになる．この場合，いずれの方法によっても高度に有意な結果が得られるが，有意確率は手法によって随分異なることが分かる．max χ^2 が極めて高度に有意となるのは両用量の相対有効率がカテゴリ 4，5 の間で段差的に変化しているためである．なお，max χ^2 は臨床試験の分野とは別に 2 項比率の時系列データの変化点解析(change point analysis) でも用いられている（Worseley, 1986）．

表 4　表 3 に各種検定を適用した結果

	統計量	両側有意確率
Wilcoxon：	2.488	0.0128
χ^{*2} ：	18.453	0.0096
max $\chi^2=\chi^2$ (1,2,3,4；5.6)：	10.303	0.0033

順序分類データを扱うもう一つの方法は背景に連続分布を想定し，処理効果の差違を背景分布の位置あるいは尺度母数の差違で説明することである．このアプローチは古く，たとえば官能検査データにロジスティック分布を当てはめた Snell (1964) にも見られるが，日本でよく用いられるようになったのは McCullagh (1980) および McCullagh & Nelder (1983, 86) 以降である．よく用いられるのは処理 i ごとに位置母数の異なるロジスティック分布

$$F(x)=\frac{\exp(\theta_i+\beta x)}{1+\exp(\theta_i+\beta x)}$$

を想定し，順序分類データが未知ではあるが処理 i に依らない共通の区分点 x_j で区切られた頻度データとして観測されたと仮定することである．これは第 i 処理の第 j カテゴリの生起確率を p_{ij} とするとき，j カテゴリまでの累積確率

$$P_{ij}=p_{i1}+\cdots+p_{ij}$$

に対し

$$\log\frac{P_{ij}}{1-P_{ij}}=\theta_i+\beta x_j \tag{4.1}$$

というモデルを想定することに帰着する．モデル (4.1) は比例オッズモデルと呼ばれる．処理比較に当たって $\tau_j=\beta x_j$ は未知の攪乱母数であることに注意する．

尺度母数のモデ化には比例ハザードモデル

$$\lambda(x)=\exp(\theta_i)\lambda_0(x) \tag{4.2}$$

から導かれる

$$\log\{-\log(1-P_{ij})\}=\theta_i+\log\int_0^{x_j}\lambda_0(x)dx \tag{4.3}$$

がよく用いられる．$\tau_j=\log\int_0^{x_j}\lambda_0(x)dx$ は撹乱母数である．モデル (4.3) も比例ハザードモデルと呼ばれるがその名はハザード関数 $\lambda(\chi)$ が基準ハザード関数 $\lambda_0(\chi)$ の定数倍というモデル (4.2) に由来している．

この種のモデルは他にもいろいろあり，たとえばMcCullagh (1980) では位置，尺度母数ともに異なるロジスティック分布を想定することによりSnell (1964) のあてはめを改良している．また，行単位，列単位の多重比較からブロック交互作用モデルをあてはめる方式の提案もある (Hirotsu, 1983)．

5. 共変量解析

序論で述べたように，臨床試験には年齢，性別，重症度，起因菌，地域など被験者の不均一性の他，施設など予後に有意な影響を与えるかもしれない多くの要因が存在する．これらの変数は共変量 (covariate) と呼ばれ，この影響度を推定し，処理比較からそれを除去することが行われる．このような共変量解析は，完全無作為化の原理に基づき共変量をすべてノイズとみなすdesign basedの解析に対してmodel basedの解析と呼ばれる (Koch & Edwards, 1985)．design basedな解析の結果は無作為標本の抽出された当該母集団に限って適用されるのに対し，model basedな解析結果はより一般の母集団に適用され得るが，そのためにはモデルの適合度チェックが十分に行われなければならない．次章で述べる分割表タイプの層別解析も古くから行われている共変量解析の一種であるが，共変量解析の名が定着し，実際にもよく用いられるようになったのはCox (1972) の生存時間解析以来である．前章で述べたMuCullagh (1980) およびMuCullagh & Nelder (1983, 86) の一般線形モデルもその流れを継承したものである．

Cox (1972) は生存時間解析において共変量をとりこんだ比例ハザードモデル

$$\lambda(t;\boldsymbol{z})=\exp(\boldsymbol{\beta}'\boldsymbol{z})\lambda_0(t) \tag{5.1}$$

を想定した．前章のモデル (4.2) は $\boldsymbol{z}$ がベクトルダミー変数の場合に当たっている．このモデルは生存時間関数 F について基準生存時間関数のべき乗

$$F(t)=\{F_0(t)\}^{\exp(\boldsymbol{\beta}'\boldsymbol{z})} \tag{5.2}$$

(いわゆる Lehmann alternative) を想定するのと同値である．

モデル (5.1), (5.2) のように，分布関数は未知関数としたまま共変量効果の影響をモデル化する方式は，このような構造をまったく仮定しないノンパラメントリックな方式に対してセミパラメトリックモデルと呼ばれる．たとえば，いわゆる 2 標本問題（単純な 2 処理比較）において分布形を仮定せずに単に大小関係を論ずるのはノンパラメトリックなアプローチだし，ロケーション分布族の位置母数変化として定式化すればそれはセミパラメトリックなアプローチになる．セミパラメトリックなアプローチは Cox (1972) 以降系統的なアプローチがなされるようになったが，とくにモデル (5.1) は Cox リグレッションの名でも呼ばれる．

Cox は比例ハザードモデル (5.1) の解析において，未知攪乱関数 (nuisance function) $\lambda_0(t)$ の消去と生存時間解析で不可避な右側打ち切りに対処するため，被験者 i の死亡時点 t_i における条件付死亡率を，t_i の直前のリスク集合 R_i を用いて

$$\frac{\lambda(t_i;\boldsymbol{z}_i)}{\sum\limits_{k\in R_i}\lambda(t_i;\boldsymbol{z}_k)}$$

と定義し，部分尤度

$$L(\boldsymbol{\beta}]=\frac{\exp(\boldsymbol{\beta}'\boldsymbol{z}_i)}{\sum\exp(\boldsymbol{\beta}'\boldsymbol{z}_k)}$$

に基づく議論を展開した．このアプローチは，$\boldsymbol{z}_i$ がスカラーで 2 値をとる 2 標本問題で，タイおよび打ち切りがなければログランクスコア (3.1) を導く．しかしながらより一般の場合に $L(\boldsymbol{\beta})$ の尤度としての合理性を示し，それに基づく最尤推定量の最適性を示したのは後の論文 (Cox, 1975) である．

AIDS の抗ウイルス剤開発においても共変量解析は重要な手法であるが，いくつか AIDS 研究特有の難点を有する．最近の AIDS 研究における主たる興味は，HIV (Human Immuno Deficiency) 感染後 AIDS 発症までの潜伏期間延長に関する処理効果比較と将来の各時点での AIDS 患者数予測である．このうち前者は一種の生存時間解析であるが，次のような固有の問題を有する．データとしてはいくつかの観測時点までの HIV 感染者が左側打切データ (一般に正確な感染時点が不明) として得られる．その際，通常は観測時点で既に AIDS を発症している患者が除外されるが，それはより潜伏期間の長い患者をより多くサンプリングすることによる潜伏期間の過大評価をもたらす．一方輸血による感染の場合に AIDS 発症から遡って輸血時点を感染時点としてサンプルに算入すると，未発症者が除外されるための潜伏期間過小評価が生じる．次に潜伏期間の分布が求められたとして，AIDS 発症者の区間データから将来の患者数を予測するには，感染の時点分布と現時点までの感染者総数の推定を要する．このうち前者について，未知の感染時点を X，潜伏期間を T，AIDS 発症時点を Z とし，

$$X+T=Z$$

の関係から T, Z に関する情報を基に deconvolution によって X の分布を推定する後退法 (back-calculation method) が提案されている．ただし，その際 X と T の独立性は仮定されている．後者については現時点までに AIDS を発症していない感染者数を推定するいわゆる大きさの推定 (size estimation) の問題が生じ，統計的には極めて ill-posed な問題となる．さらに，基となる AIDS 発症者数の区間データに関し，病気の性質上過少報告 (under reporting) の問題があり，これは予測に過小評価をもたらす．また，現時点と予測時点の間の新たな感染者を予測にどう含めるか，診断の精密化，基準の変更等現在進行中の要因の影響を予測にどう採り入れるか，HIV 不感症の異質集団の存在を仮定する場合その割合をどう推定するか等の問題がある．これらについてより詳しくは Jewell (1990) のサーベイを参照されたい．

共変量を扱うもう一つの流れは2項比率データに対するロジットモデルである．本来これはバイオアッセーの分野で毒物の量 x に対する死亡率 p_x のモデル化から発したものである．この場合，p の範囲 $0<p<1$ を無視した回帰モデルが不適切なのは明白である．そこで当該動物母集団の毒物に対する閾値分布関数（ロジスティック分布）$F(x)$ を用いて

$$F(x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} \tag{5.3}$$

とすると，ロジット $\log\{p/1-p)\}$ に関する回帰モデル

$$\log\frac{p}{1-p}=\beta_0+\beta_1 x \tag{5.4}$$

が得られる．ロジットが $0<p<1$ を $\pm\infty$ の範囲に拡げている点が重要である．式(5.4)は式(4.1)と形式上類似しているが，比例オッズモデルでは x（区分点）が未知攪乱母数であるのに対し，ロジットモデルでは x が既知の説明変数であることに注意して欲しい．これ故に，ロジットモデルはこのような閾値モデルに限らず，一般に2項比率のモデル化に用いられる．すなわち共変量ベクトルを $\boldsymbol{x}$ として

$$\log\frac{p}{1-p}=\boldsymbol{\beta}'\boldsymbol{x}$$

とモデル化することがよく行われ，ロジスティック回帰モデルと呼ばれている (Koch & Gillings, 1983)．

6. 分割表解析

重症度や急性，慢性のような共変量は層別因子として扱われることが多く回帰モデルとしてより，3次元分割表として定式化されることが多い．この場合注意すべきことは

（1）　層別因子の水準によって処理効果の差に違いがあるか否か，

（2）　処理効果の差に違いがないとき，層別因子の水準を通しての処理効果の差をどう定義すればよいか

の2点である．これを実例に基づいて説明する．

表4は複雑性尿路感染症に関する二重盲検比較試験の結果である．この例では腎盂腎炎で治験薬と対照薬の効果に逆転が見られる．そのために併合した表では効果が相殺し，p-値は 0.393 と大きな値になる．もし層による違いが偶然のばらつきを超えて有意なものなら併合表での推論は医学的にも数学的にもミスリーディングであることは明白である．このように層により二つの薬の相対的な薬効差が異なることを層と薬の交互作用と呼ぶ．

表4　複雑性尿路感染症二重盲検比較試験

層別 H	薬剤 I	改善の有無 J 1.無効	2.有効	例数	有効率	p-値
1. 腎盂腎炎	1. Treatment	7	14	21	0.677	0.235
	2. Control	3	15	18	0.833	
2. 膀胱炎	1. Treatment	13	62	75	0.827	0.128
	2. Control	18	46	74	0.719	
計	1. Treatment	20	76	96	0.792	0.393
	2. Control	21	61	82	0.744	

次に膀胱炎二重盲検比較試験の結果（表5）について考える．表5では単純性，複雑性ともに顆粒が優れており，上に述べたような交互作用は見られない．しかるに併合した表でのp-値は各層でのp-値に比べかえって増大している．本来，似たような証拠を二つ合わせれば証拠は強まるはずであり，この場合も併合はミスリーディングである．この例では単純性と複雑性で有効率が相当異なり，カプセル剤が有効率の高い単純性でより多く集められたため，単なる併合がカプセル剤に有利に働いたのである．この例を煎じつめると層別した各表では第一薬が優るのに併合した表では第二薬が優るということも起こり得，Simpsonのパラドックスと呼ばれている．

表5　膀胱炎二重盲検比較試験

層別 H	薬剤形 I	改善の有無 J 1.無効	2.有効	例数	有効率	p-値
1. 単純性	1. カプセル	15	82	97	0.845	0.092
	2. 顆　粒	2	38	40	0.950	
2. 複雑性	1. カプセル	30	23	53	0.434	0.073
	2. 顆　粒	16	26	42	0.019	
計	1. カプセル	45	105	150	0.700	0.188
	2. 顆　粒	18	64	82	0.781	

さて，層別解析では上に述べた二つの現象，すなわち層と薬剤の交互作用およびSimpsonのパラドックスに気を付ける必要がある．これに対処する統計的方法としては交互作用検定のためのBreslow-Day検定および交互作用がないときに，Simpsonのパラドックスを避けて薬効差を比較するためのMantel-Haenszelの検定がよく知られている．Mantel-Haenszelの方法は併合した表で薬効差を測る代わりに，いわば層ごとに薬効差を測りそれを併合するものである．Breslow-Day検定およびMantel-Haenszel検定は現実によく用いられるが，これと漸近的に等価で数学的により明解なのは対数線形模型によるアプローチである．

いま，層 i において薬剤 j が第 k カテゴリに反応する確率を p_{ijk} と表し，対数線形模型

$$\log p_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk}$$

を想定する．セル度数を y_{ijk} とするとき，確率分布としては $y_{...}$ のみ与えた多項分布（$p_{...}=1$），あるいは $y_{ij.}$ を与えた独立な多項分布の積（$p_{ij.}=1$）などいろいろ考えられるが，仮説検定のさいに十分統計量を与えた条件付推測の立場をとるならこれらカーネルの等しい確率分布はすべ

て同値な結果を与える．さらにそれは1次の漸近論では無条件検定とも一致する．そこで分かり易いのはむしろ y_{ijk} に独立な poisson 分布 $P(m_{ijk})$ を仮定し，m_{ijk} に関する対数線形模型

$$\log m_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk} \qquad (6.1)$$

を想定することである．層と薬剤の交互作用に関する帰無仮説はこのモデルで

$$H_{\alpha\beta\gamma}:(\alpha\beta\gamma)_{ijk}=0$$

と表される．$H_{\alpha\beta\gamma}$ が成り立たなければ層を併合する意味がなく，層ごとに異なった解釈をすることになる．$H_{\alpha\beta\gamma}$ が成り立つとき，$(\beta\gamma)_{jk}$ が層を通しての薬効差を表すが，$(\alpha\beta)_{ij}$ と $(\alpha\gamma)_{ik}$ が共に存在すると併合表による $(\beta\gamma)_{jk}$ の評価がミスリーディングになる．$(\alpha\beta)_{ij}$ は層における各薬剤例数のアンバランス，$(\alpha\gamma)_{ik}$ は層による有効率の違いを表す．$H_{\alpha\beta}:(\alpha\beta)_{ij}=0$, $H_{\alpha\gamma}:(\alpha\gamma)_{ik}=0$，のいずれか一方でも成立すれば Simpson のパラドックスは生ぜず併合表での薬効評価が正当化される．そこで実際的な手順は Poisson 分布モデル（6.1）において一連の仮説 $H_{\alpha\beta\gamma}$, $H_{\alpha\beta}$, $H_{\alpha\gamma}$ そして $H_{\beta\gamma}:(\beta\gamma)_{jk}=0$ の尤度比検定を行うことである．$H_{\alpha\beta\gamma}$ の検定が Breslow-Day 検定，$H_{\alpha\beta\gamma}$ のみを仮定した $H_{\beta\gamma}$ の検定が Mantel-Haenszel 検定に対応し，$H_{\alpha\beta\gamma}$ ならびに $H_{\alpha\beta}$, $H_{\alpha\gamma}$ の少なくとも一方を仮定した $H_{\beta\gamma}$ の検定は層を併合した2元表での独立性検定（交互作用検定）に帰着する．この手順を適用すると表4では $H_{\alpha\beta\gamma}$ の検定が有意確率 0.074 となって傾向差を示す．これはモデル検定としては有意と見なしてよいだろう．表5では $H_{\alpha\beta\gamma}$ の有意確率は 0.564 だが，$H_{\alpha\rho}$ と $H_{\alpha\gamma}$ の有意確率はそれぞれ 0.002, 0.000 となり Simpson のパラドックスの生じる状況であることが分かる．$H_{\alpha\beta\gamma}$ のみを仮定した $H_{\beta\gamma}$ の検定は有意確率 0.013 を与え，併合した表に対する結果（有意確率 0.188）に比べ合理的な結果を与える．

この対数線形模型によるアプローチは層および薬剤が多水準の場合にも容易に拡張される．また，検定の各ステップで最尤推定量を求めるのに，帰無仮説の下での十分統計量（周辺和）を与えた比例反復法 (Fienberg, 1980) が知られている．なお，分割表データの共変量解析については Koch et al. (1982) のレビューが参考になる．

7. 多重比較法

表1は実は下記表6の一部であり，2群比較としての取扱いは適切ではない．

表6　NFLX 半減期（ラット）

投与量 i (mg/kg/day)		繰り返し j				
1.	5	1.17	1.12	1.07	0.98	1.04
2.	10	1.00	1.21	1.24	1.14	1.34
3.	25	1.55	1.63	1.49	1.53	2.14
4.	50	1.21	1.63	1.37	1.50	1.81
5.	200	1.78	1.93	1.80	2.07	1.70

この例では水準に自然な順序があり，投与量に応じた半減期の増大に興味がある．すなわち半減期の一様性の帰無仮説をいわゆる傾向性仮説に対して検定することに興味が持たれる．正規分布モデル $N(\mu_i, \sigma^2)$, $i=1,\cdots,5$, を仮定すればそれは

帰無仮説　$H_0:\mu_1=\cdots=\mu_5$

を

対立仮説　$H_1:\mu_1\leq\cdots\leq\mu_5$（少なくとも一つの不等号は厳密）　(7.1)

に対して検定する問題として定式化される．式（7.1）は丁度2項比率に対する式（3.2）に対応しており，線形統計量によるトレンドテストや累積 χ^2 法などが適用される（広津，1992a）．しかしながら実際上必要なのはそのような帰無仮説の一様性検定ではなく，零レベルと有意に異なる応答を示すレベルや，前後で大きな段差を示すレベルを検出することである．前者の目的では Williams（1971）の多重比較法および Marcus（1976）によるその修正，そして後者の目的では Hirotsu et al.（1992）の max t 法が提案されている．

この他臨床試験でよく用いられるのは，複数の治験薬と対照薬を比較する Dunnett の方法である．これは μ_1 を対照薬の平均とするとき

$$\text{帰無仮説 } H_0: \mu_i=\mu_1, \quad i=2,\cdots, a$$

を t 統計量

$$t_i=(\bar{y}_i-\bar{y}_1)/\sqrt{(n_i^{-1}+n_1^{-1})\hat{\sigma}^2}, \quad i=2,\cdots, a$$

を用いて多重検定することにより

$$\text{対立仮説 } \quad H_1: \mu_i\geq\mu_1, \qquad i=2,\cdots, a$$

の各方向で検出力を高めるものである．ただし，n_i, $\bar{y}_i$ はそれぞれ第 i 薬剤の繰り返し数および標本平均，$\hat{\sigma}^2$ は自由度 $\sum_i(n_i-1)$ の群内不偏分散である．なお，すべての対比較を行う Tukey 法およびあらゆる対比の多重比較を行う Scheffe 法は水準に特別な構造を仮定しないオムニバス方式であるため，臨床試験で用いられることは比較的少ない．

多重比較法では一般に複数の統計量の最大成分を用いることから，有意確率の計算に多重積分が要求される．たとえば Dunnett 法や max t 法ではナイーブに考えて，$\hat{\sigma}^2$ に関する積分を除いて a-1 重積分が必要である．しかしながら Dunnett 法については

$$u_i=(\bar{y}_i-\bar{y}_1)/\sqrt{(n_i^{-1}+n_1^{-1})\sigma^2}$$

の相関構造が $\lambda_i=n_i/(n_i+n_1)$ により

$$\mathrm{Cov}(u_i, u_{i'})=\sqrt{\lambda_i}\cdot\sqrt{\lambda_{i'}}$$

という積で表されることから a がいくつであっても，$\hat{\sigma}^2$ および正規分布関数で表される部分を除いて1次の積分ですますことができる（Hochberg & Tamhane, 1987）．max t 法についてはその成分

$$u_I=\left(\frac{Y_I}{N_I}-\frac{\bar{Y}_I}{\bar{N}_I}\right)\Big/\sqrt{\left(\frac{1}{N_I}+\frac{1}{\bar{N}_I}\right)\hat{\sigma}^2},$$

$$Y_I=\sum_1^I n_i\bar{y}_i, \qquad \bar{Y}_I=\sum_{I+1}^a n_i\bar{y}_i,$$

$$N_I=\sum_1^I n_i, \qquad \bar{N}_I=\sum_{I+1}^a n_i$$

の相関構造が $\lambda_I=N_I/\bar{N}_I$ により

$$\mathrm{Cov}(u_I, u_{I'})=\sqrt{\lambda_I}/\sqrt{\lambda_{I'}}, \quad I<I' \tag{7.2}$$

という商の形で表されることから，$\hat{\sigma}^2$ および正規分布関数で表される部分を除いて $[\log_2(a-1)]$ 次の積分でよいことが分かる（栗木他，1989）．なお，正規分布に関する式（7.2）のような相関構造が u_I の両側マルコフ性を意味し，またその逆も成り立つことに注意する．

順序分類データで多重比較法が必要とされる場合も多い．たとえば表3も実は表7の一部であり，用量の順序を考慮した多重比較法の適用が望まれる．

表7　用量設定試験における全般有用度

全般有用度 j 薬剤 i	1. 好ましくない	2. やや好ましくない	3. どちらともいえない	4. やや有用	5. 有用	6. 極めて有用	計
1. Af6mg	5	6	21	16	23	6	$77=n_1$
2. Af 3 mg	7	4	33	21	10	1	$76=n_2$
3. Placebo	3	6	37	9	15	1	$71=n_3$

この場合は用量設定試験なのでAF 6とAF 3以下またはAF 3以上とプラセボ（零レベル）のどちらに有意な段差があるかを見るための max t 型の多重比較法が適当である．ただし，3，4章で述べたように順序応答に関しては線形スコア，累積 χ^2 そして max χ^2 型の統計量の選択の余地がある．この例にこれらの統計量に基づく max t 型の多重比較法を適用すると，それぞれ両側有意確率 0.011，0.005，0.014 でAF 6とAF 3以下の間に有意な段差が示唆される．

次に表8では第一薬剤が対照薬であり，Dunnett 法の適用が適切である．実際，表8に Kruskal-Wallis 法を適用すると有意確率 0.07 であるのに対し，同じ Wilcoxon スコアを用いた Dunnett 法では有意確率 0.048 となって検出力が上がることが観察される．このように多群比較においては焦点の定まらない一様性検定に替えて目的に応じた多重比較法の適用が薦められる．

表8　抗生物質比較臨床試験

全般有用度 j 薬剤 i	1. 無効	2. やや有効	3. 有効	4. 著効	計
1. AMPC	3	8	30	22	$63=n_1$
2. S6742	8	9	29	11	$57=n_2$
3. CCL	2	11	33	17	$63=n_3$

8. 経時測定データの解析

臨床試験では1時間おきに測定した血中濃度，4週おきに6期測定した体内総コレステロール量，1カ月おきに1年間測定した血圧値など，一人の被験者につき経時的な一組のデータが得られることが多い．これは一種の時系列データであるが，比較的短期時系列であること，およびこれらのデータに基づく処理比較が目的となることが多いことから多変量モデルのアプローチが採られることが多い．すなわち，第 h 処理を施された第 i 被験者の t 期にわたる測定値ベクトル $y_{hi}=(y_{hi1},\cdots,y_{hit})'$ を

$$\boldsymbol{y}_{hi}=\boldsymbol{\mu}_h+\boldsymbol{\varepsilon}_{hi}$$

と表し，$\boldsymbol{\mu}_h$ は未知定数ベクトル，誤差ベクトル $\boldsymbol{\varepsilon}_{hi}$ は互いに独立に正規分布 $N(\boldsymbol{0},\Omega)$ に従うことを仮定する．とくに経時パターンを時間に関する多項式で説明するモデルは成長曲線モデルと呼ばれ，層別因子による成長曲線の違いを検証するのに一般化多変量解析の手法が適用される．これらのモデルは被験者を単に繰り返しと見ていること，処理効果は平均のみに影響を与えること，そしてデータの共分散を経時相関として捉えていることに特徴がある．共分散構造としてはパラメータ節約のため AR モデル（Ware, 1985）やスフェリカルな構造

$$\Omega=\sigma_w^2 I+\sigma_b^2 \boldsymbol{j}\boldsymbol{j}' \quad (I=\text{単位行列},\ \boldsymbol{j}=(1,\cdots,1)') \tag{8.1}$$

を仮定することが多い．式（8.1）のもとでは標準的な分割法（split plot design）データの解析が適用される．しかしながら，ブロック内での無作為化を前提とする分割法と，順序に意味のある経時データは本質的に異なり，このようなスフェリカルな構造が適合することはむしろ少ない．

経時的なプロファイルをより積極的に分散構造のモデル化に応用したものとしてはLaird & Ware（1982）の2段階モデルがある．今最も簡単な，時間xに関する単回帰モデル

$$y_{hij}=\alpha_h+\beta_h x_j+\varepsilon_{hij} \tag{8.2}$$

を考える．ここで被験者iのばらつきを考慮し$\beta_h \to \beta_h+b_{hi}$とおきかえる．ただし，$b_{hi}$は被験者の母集団上で期待値0，分散$\sigma_b^2$で，たがいに無相関な確率変数とする．この設定では誤差ε_{hij}はたがいに無相関で，かつb_{hi}とも無相関と仮定して無理がない．するとy_{hij}の相関構造として

$$\mathrm{Cov}(y_{hij}, y_{h'i'j'})=\delta_{hh'}\delta_{ii'}(\sigma_b^2 x_j x_{j'}+\sigma^2\delta_{jj'})$$

が導かれる．つまりこのモデルではデータの相関構造は系列相関というより，被験者の応答の不均一性から生じると仮定している．このモデルは上で述べた考え方のプロセスに従って混合模型または2段階モデルと呼ばれ，元の回帰モデル（8.2）の構造によっていろいろ複雑な分散構造をモデル化することができる．また，回帰モデルを仮定することにより，すべての被験者につき測定時点が一定でないいわゆるアンバランストケースも取り扱うことができる．2段階モデルにつき詳しいこと，および最尤推定量を求めるアルゴリズムについてはCrowder & Hand（1990）を参照されたい．

とくに薬剤動態学（pharmacokinetics）の分野では患者内および患者間の時間変動を表すのに非線形モデルが本質的である．そこで上に述べた2段階線形モデルを拡張したNONMEM（Nonlinear Mixed Effects Model）をはじめ様々な方法が提案されている．その計算アルゴリズムや最近の研究のサーベイについてはVonesh & Carter（1992）を参照されたい．

被験者の応答の不均一性を積極的に認め，標示因子として扱ったものとして広津（1989）およびHirotsu（1991）がある．この方法はまず被験者のプロファイルの非類似性を表す累積χ^2統計量の多重比較により，被験者を群分けする．累積χ^2が時間軸に沿った上昇，下降のような系統的乖離をよく検出するため，各群は通常，臨床的意味の明確な特徴付けがなされる．そこで各処理に属する被験者が各群にどのように分布するかによって処理を特徴付けることができる．丹後（1989）ではあらかじめ群を特徴付けるいくつかの回帰パターンを準備しておき，各被験者をそのうちどのパターンに最もよく適合するかによって分類する方式が提案されている．これらはいずれもある一定の処理に対し，偶然変動の範囲を超えて良く反応する被験者もいれば悪く反応する被験者もいるという実臨床上の経験をモデル化したものである．

9. 同等性検証

同等性検証はアクティブな対照薬を用いる臨床試験に固有，かつ本質的な問題である．現在，治験薬は対照薬に対し，有効率において必ずしも有意に優れている必要はなく，同等であればよいと考えられている．それは次のような理由によっている．

（1） 薬剤の特性は一面的ではなく，単なる薬効のほかに，副作用や投与法の容易さなどいろいろな側面がある．たとえばマイルドで副作用の少ない治験薬が対照薬と同等の臨床効果を持てばその薬は有用と考えられる．また，1日3回投与が1回ですみ，しかも同等の薬効が期

待できるという薬剤が開発されたならそれもまた有用である．

（2） 同等の薬が共存し，競合することは追跡調査，改良などの意欲を促し，あるいは薬価の面でよい影響をもたらす．

以上の観点から，同等なら認可という考え方そのものは至極自然に思われる．問題は同等性検証のための適切な統計的方法がなく，長く有意差検証のための検定が行われ，有意水準5％で有意でないことをもって同等とみなしてきたところにある．そのため有効率で10％内外劣っている可能性の否定できない薬が実際に認可されてきている（広津，1992b）．さらにこの方式の難点は，効率の高い，良質の臨床試験を行うというインセンティブそのものが失われ勝ちなことである．

考えられる一つの方法は，臨床的に許容できない平均値，あるいは有効率の差を定義し，それに対して一定の検出力を保証する例数を要求することである．しかしながらその方法では一般に実際的とはいえない膨大な例数が要求されてしまう．あまりに保守的過ぎると，良い薬をなるべく早く世に出したいという臨床試験の持つもう一つの側面にもとることになる．

そこで考えられるのは，たとえば有効率の場合に，治験薬，対照薬それぞれの有効率を p_1, p_0 として

$$\text{帰無仮説}\quad H_0: p_1 = p_0 - \Delta$$

を

$$\text{対立仮説}\quad H_1: p_1 > p_0 - \Delta \tag{9.1}$$

に対して検定し，有意差を示すことである．これだと治験薬が対照薬に対して△以上劣っていないことを積極的に証明するため，治験の質を向上させることにもなる．また，この方式で有意水準 α，検出力 $1-\beta$ としたときに必要な例数は1群当たり

$$m > \left(\frac{\rho K_\alpha + K_\beta}{p_1 + \Delta - p_0}\right)^2 \{p_0(1-p_1) + p_0(1-p_1)\},$$

$$\rho = \sqrt{2 \times \frac{p_0+p_1}{2}\left(1 - \frac{p_0+p_1}{2}\right) \Big/ \{p_0(1-p_0) + p_1(1-p_1)\}}$$

となる（広津，1992b）．従って治験薬が対照薬に対し真に同等以上だと有意差を示すのに極めて少ない例数ですむという，いい意味でのめりはりが出てくる．なお，式(9.1)の片側検定方式で有意となることは，有効率差 $p_1 - p_0$ に関し信頼率 $1-2\alpha$ の信頼区間を構成し，その左端が $-\Delta$ を超えることと同値である．

実際に Δ をどう選ぶかという問題があるが，希用薬などは除いた一般の場合で $p_0 = 0.2 \sim 0.8$ の範囲に対しとりあえずの基準として $\Delta = 0.1$ が提案されている．すべてをこの基準にあてはめる必要もないが，一方，疾患ごとにきめ細かく△を変えるというのも実際的ではない．また，有効率に替えてロジット差で規制することも考えられるが，上記の有効率の範囲で両者が大きく異なることがないなら，ナイーブな有効率差の方が臨床的解釈が仕易いだろう．

新薬が単なる剤形変更や，投与経路変更のときには，臨床試験を一からやり直すことなしに，血中濃度などの同等性を証明すればよいことになっている．これを生物学的同等性の証明という．この場合，統計モデルとしては正規分布を用い，生物学的ばらつきを考慮して平均の相対差 $|\mu_1 - \mu_0|/\mu_0$ が20％以内であればよいとの基準が示されている（江島他，1982）．これについても有効率差の場合と同様の考え方が適用できるが，相対差を厳密に取り扱う統計的方法は今後の研究にまつところが大きい．

10. おわりに

臨床試験に関してはここに書ききれなかった問題が他にも数多く存在する．たとえば，一つ

の新薬をきちんと評価するのに膨大な時間と人手と費用を要するため，各国で行われた治験を相互に利用できないかとの議論がある．しかし，このことは本文中で述べた地域，人種差に関する被験者母集団の representativeness とはある意味で矛盾する．

また，抗生剤の領域で日米の根本的な思想の相違として議論されている問題がある．すなわち米国では救命の観点から殺菌効果を重視し，概して高用量が採られる．一方，我国では副作用を重視し，できるだけ低い用量で菌の生育を押さえ，患者自身の防御機構で治癒に至ることを期待している．従って，それぞれに薬剤の'有用性'を評価すると結果に大きな食い違いが生じる可能性がある．さらに，米国では疾患ごとに対象起因菌を特定しているのに対し，日本ではそのような組み合わせ規制は採っていない(島田，八木澤，1992)．この場合，米国流が厳密に見えるものの，実際の臨床現場では菌の同定前に投薬が行われることも多々あるとすれば，一概に日本流が無意味ともいえない．

このように統計解析以前に，治療に関する根本的な見解の統一，また，より科学的側面を重視するのか，より技術的側面を重視するのかといった視点の統一にも多くの議論を費やす必要がある．

謝辞：査読者から初稿に対しいくつかの貴重な意見を頂戴した．それによって追加した項目もあることを記して謝辞に替えたい．

参 考 文 献

[1] Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist* **23**, 193-212.
[2] Buyse, M. E., Staquet, M. J. and Sylvester, R. J. (eds.) (1988). Cancer clinical trials : methods and practice. Oxford : Oxford University Press.
[3] Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist.* B **74**, 187-220.
[4] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.
[5] Cox, D. R. (1982). A remark on randomization in clinical trials. *Utilitas Mathematica* **21** A, 245-252.
[6] Crowder, M. J. & Hand, D. J. (1990). Analysis of repeated mesures. London : Chapman and Hall.
[7] Efron, E. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58**, 403-417.
[8] 江島昭也 (1982). 生物学的同等性の試験方法についての解説. 医薬品研究 **13** (5), 1106-1119.
[9] Fienberg, S. E. (1980). The Analysis of cross-classified categorical data. (2nd ed.). Cambridge, MA : MIT Press.
[10] Hirotsu, C. (1979). An F approximation and its application. *Biometrika* **66**, 577-584.
[11] Hirotsu, C. (1983). Defining the pattern of association in two-way contingency tables. *Biometrika* **70**, 579-589.
[12] Hirotsu, C. (1986). Cumulative chi-squared statistic as a tool for testing goodness of fit. *Biometrika* **73**, 165-173.
[13] 広津千尋 (1989). 経時測定データ解析のためのモデルとその応用. 日本品質管理学会誌 **19**, No. 3, 16-23.
[14] Hirotsu, C. (1991). An approach to comparing treatments based on repeated mesures. *Biometrika* **78**, 583-594.
[15] 広津千尋 (1992a). 実験データの解析—分散分析を超えて—. 東京：共立出版.
[16] 広津千尋 (1992b). 臨床試験データの統計解析. 東京：広川書店.
[17] Hirotsu, C., Kuriki, S. & Hayter, A. J. (1992). Multiple comparison procedures based on the maximal component of the cumulative chi-squared statistic. *Biometrika* **79**, 381-392.
[18] Hochberg, Y. & Tamhane, A. C. (1987). Multiple comparison procedures. New York : John Wiley & Sons.
[19] Jewell, N. P. (1990). Some statistical issues in studies of the epidemiology of AIDS. *Statistics in*

Medicine **9**, 1387-1416.

[20] Koch, G. G., Amara, I. A., Davis, G. W. & Gillings, D. B. (1982). A review of some statistical methods for covariance analysis of categorical data. *Biometrics* **38**, 563-595.

[21] Koch, G. L. & Edwards, S. (1985). Logistic regression. In Encyclopedia of Statistical Sciences 5, Johnson, N. V. and Kotz, S. (eds.). New York : John Wiley & Sons, 128-133.

[22] Koch, G. G. & Gillings, D. B. (1983). Inference, design based vs. model based. In Encyclopedia of Statistical Sciences 4, Johnson, N. L. and Kotz, S. (eds.). New York : John Wiley & Sons, 84-88.

[23] 厚生省（1992）．臨床試験の統計解析に関するガイドライン．

[24] 栗木 哲，広津千尋，Hayter, A. J.（1989）．累積カイ二乗の最大成分に基づく多重比較法—有意確率計算と用量水準比較への応用—．応用統計学 **18**, No. 3, 129-141.

[25] Laird, N. M. & Ware, J. H. (1982). Random-effects model for longitudinal data. *Biomerics* **47**, 1557-1561.

[26] Marcus, R. (1976). The powers of some tests of the equality of normal means against an ordered alternative. *Biometrika* **63**, 177-183.

[27] McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc.* B **42**, 109-142.

[28] McCullagh, P. & Nelder, J. A. (1983). Generalized linear models. New York : Chapman and Hall.

[29] McCullagh, P. & Nelder, J. A. (1986). Generalized linear models. (2nd ed.). New York : Chapman and Hall.

[30] Meinert, C. L. (1986). Clinical trials : design, conduct, and analysis. New York : Oxford University Press.

[31] 尾上能之，広津千尋，栗木 哲(1992)．いくつかのノンパラメトリック検定の正確な有意確率評価．日本応用統計学会年会予稿集，8-12.

[32] Pocock, S. J. (1983). Clinical trials. New York : John Wiley & Sons.

[33] 島田 馨，八木澤守正（1992）．抗菌剤の日米欧の比較．化学療法の領域 **8**, 1732-1740.

[34] Snell, E. J. (1964). A scaling procedure for orderd categorical data. *Biometrics* **20**, 592-607.

[35] 丹後俊郎(1989)．臨床試験における経時測定データの解析のための混合分布モデル．応用統計学 **18**, 143-161.

[36] Vonesh, E. F. & Cater, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* **48**, 1-17.

[37] Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *Amer. statist.* **39**, 95-101.

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

[39] Worseley, K. J. (1986). Confidence regions and tests for a changepoint in a sequence of exponential family random variables. *Biometrika* **73**, 91-104.