Clin. Eval. 14 : 467-475, 1986. Sep.

総　説

臨床試験における統計的諸問題 (1)

——同等性検定を中心として——

広津千尋*

Some Statistical Problems in Clinical Trials (1)

——Test for the Equivalence of Two Drugs——

Key words: Confidence interval; Negative trial; Test for equivalence

Chihiro Hirotsu: Faculty of Engineering, University of Tokyo.

Non significance by the usual test of the null hypothesis gives by no means an evidence for the equivalence of two drugs without any restriction on the sample size. Therefore a method is described to prove that a new drug is inferior to a control by no more than a prescribed constant. The method is applied to some recent clinical trials in Japan to illustrate some of the ideas.

The main topic in the subsequent paper is the multiplicity in clinical trials.

* 東京大学工学部計数工学科

1. 序論

本稿はここ数年臨床試験の統計解析に関与した経験から現在普通に行われている解析法に関していろいろな問題を指摘し，その解決策を探ろうとするものである．それらは単なる誤用から厳密な科学的議論を要するもの，さらに相当な議論を費してもなお明快な結論の得られないものなどレベルはまちまちである．

その中で最大の問題点はやはり治験薬と対照薬の同等性の検証であろう．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．それはおそらく次のような理由によるものであろう．

(1) 薬剤の特性は一面的ではなく，単なる薬効の他に，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だろう．また，1日3回投与が1日1回投与ですみ，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，それもまた有用だろう．

(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持たらすだろう．

このような観点から，「同等なら認可」という考え方自身は至極もっともに思われる．問題は「同等性検証」のための統計的方法として「有意差検証」のための検定が行われ，有意差が検出されなかったことをもって同等性がいえたと見なす現状にある．いうまでもないことだが，有意水準αの有意差検定で同等性の帰無仮説が棄却されたなら，それは積極的に（危険率αで）薬剤間の有意差を示唆するが，棄却されないことは単に有意差を示す十分な証拠がないということであって，積極的に同等性を主張するものではない．いいかえると同等なら滅多に起こらないことが生じたら有意差ありということであって，生じないからといって同等ということにはならない．ここに論理のすりかえがあることは多くの人が指摘するところである．とくに有意水準αは伝統的に0.05に採られることが多いが，これは薬剤間に差がなければ100回中5回（片側でいえば2.5回）も生じないような統計値の差を規準として有意差を認める大変保守的な検定方式である．しかも現在は検出力に関する要請を課していないため，通常見られるサンプル・サイズではよほど治験薬が劣っていないかぎり同等と見なされてしまう可能性が強い．そもそもサンプル・サイズの小さな不確かな臨床試験を行うほど劣った治験薬が通りやすいというのはちょっと考えても不合理だろう．さらに，一度劣った治験薬が認可されるとこれを対照薬としてさらに劣った治験薬が認可されるというおそれも否定できない．

「同等性検証」のための統計的方法としてはいくつかの提案があるが，いずれも多少の欠点を伴っている．たとえば有効率のある程度以上の差Δを見逃す確率を一定値β以下に押えることを要求することが考えられるが，具体的にΔ，βをどのような値に設定するかという問題が生じる．また，“常識的”に$\Delta=0.1$，$\beta=0.1$と選ぶと良い治験薬にも悪い治験薬にもおしなべて相当な（現状の数倍から十数倍の）サンプル・サイズを要求することになる．本稿の第2節ではネガティブ・トライアル(negative trial)の考え方に基づき，ある程度以上劣っていないことを検証する方法について論じる．この場合，データに基づいて，差がΔ以内であることを統計的に検証するには，治験薬の有効率にΔを上乗せした上で，対照薬より有意に優れていることを危険率αでいえばよい．それには通常の「統計的有意差検定」の方式が使える．2節で詳しく述べるが，この方法によると良い薬剤の有効性をいうためには現状よりむしろ少なめのサンプル・サイズで十分だが，有効性の不確な薬剤が劣っていないことを証明するには相当なサンプル・サイズを要することになる．このことは大変合理的に思える．

この方法は本来，治験薬に副作用が少ないとか，投与法が簡便とかの利点があるときに標準薬より有効率においてΔまで劣ることを認めようという考え方に基づいてはいるが，実際には，サンプル・サイズが無限大でもないかぎり，平均出検品質（長い目で見た市場の品質）が標準薬より本当にΔ劣るということにはならない．いったん認可された後は治験薬，標準薬共に同じくらいの割合で使われるという仮定の下で，椿（東京大学工学部）の試算によれば，1群の例数が約50，$\Delta=0.1$の場合で平均出検品質の劣化は高々1～2%程度である．この意味でもこの方法は同等性検証の方法としてふさわしいものと考えられる．

同等性検証と並んで重要なのは統計的検定の多重性の問題だろう．臨床試験は本来多面性を持っている．主たる特性あるいは症状の改善度以外に様々な臨床検査値，あるいは副作用の有無が報告される．また一口に改善度といっても著明，やや有効，無効等様々な判定がありどこに注目するかという多重性があるし，経時的なものならどの時点で見るかによっても結果は変ってくる．

一方多数の背景要因が測定され，たとえば性別や重症度に関して層別解析が行われる．これらに関しては多くの議論があり（たとえば Armitage, 1984; 広津, 1984

Table 1 中間解析の例 (**Pocock** 1984から引用) (表中の数字は死亡数/例数を表わす)

		Placebo	D-penicillamine	統計値	有意確率
最初の解析	1980夏	8/32	2/55	$\chi^2=9.1$	0.003
公表結果	1981	10/32	5/55	log rank $\chi^2=7.2$	0.01
その後の成績	1984	16/37	18/61	log rank $\chi^2=3.0$	0.08

a, b; 藤田他, 1984; 藤田・椿・佐藤, 1986 など), いずれも後知恵的解析によるいい過ぎに警告を与えている. 日本では長期にわたる臨床試験の中間解析はあまり行われていないようだが, Pocock (1984b) は次のような例 (Table 1) をあげて ad hoc な統計解析を戒しめている. この例は治験実施中に何度も解析し, 有意な結果が得られたときに発表する危険性を示唆している. とくに治験計画の早期に得られた有意差から結論を急ぐのは避けなければならない. センセーショナルに効きめの報じられた新薬でその後顕著な成績の上がらないものも多いのではないだろうか.

さてこのようにたくさんある多重性の問題点の中から, 3節で2剤比較で行われる複数の検定 (3.1節) および多剤比較のための検定 (3.2節) について述べる. これらは一見違う問題に見えるが, ことの本質は同じである. 多重性の問題に同等性検証の問題がからむと話は一層複雑になる. このことについても注意を喚起したい. 層別解析については4節で述べる. 5節では経時的に採られたいわゆるリピーテッド・メジャメント (repeated measurement) の解析について簡単に述べる. この他6節では統計量の分布の近似について述べる.

2. ネガティブ・トライアルの考え方に基づく同等性検証

2.1) Δ 上乗せ方式

序論で述べたように臨床試験で通常用いられている検定方式は, いわば治験薬の有効性を検証するための方法である. したがって, 検定によって統計的有意差が検出されれば, ある危険率 α をもって治験薬の対照薬に対する優位性を主張することができる. 一方有意差が検出されないときは単に有効性が主張できないことを意味するのであって, 決して2剤間の同等性を意味するものではない.

他方, 新しく開発された治験薬がマイルド (mild) なもので, 標準薬より多少劣っていても採用する価値があると判断される場合がある. このようなときに治験薬が対照薬に対してあるあらかじめ定めた値 Δ 以上劣っていないことが危険率 α 以下で主張できたとき治験薬を採用するという考え方がある (Makuch & Simon, 1978). この場合危険率とは治験薬が対照薬より Δ 以上劣っているのに誤って治験薬を採択してしまう確率のことである. この考えに基づく臨床試験は通常の検定が"Positive"な結果を予期することに対応して "Negative Trial" と呼ばれている (Pocock, 1984a).

Table 2 臨床試験 (2剤比較) データ

	サンプル・サイズ	有効症例	無効症例
対照薬	n_0	y_0	n_0-y_0
治験薬	n_1	y_1	n_1-y_0

ネガティブ・トライアルの場合, 治験薬の有用性検証のための統計的方法は以下のようになる. すなわち, 治験薬, 対照薬の有効率をそれぞれ p_1, p_0 とするとき,

帰無仮説 $H_0(\Delta): p_1=p_0-\Delta$ (2.1)

を

対立仮説 $H_1: p_1>p_0-\Delta$

に対して有意水準 α で検定すればよい. 帰無仮説を $H_0'(\Delta): p_1\leq p_0-\Delta$ としても結果は同じになる.

いま臨床試験の結果が Table 2 のように与えられたとしよう.

このとき, 検定統計量は次で与えられる.

$$U_1=\frac{y_1/n_1+\Delta-y_0/n_0}{\sqrt{\frac{1}{n_0}\hat{p}_0(1-\hat{p}_0)+\frac{1}{n_1}(\hat{p}_0-\Delta)(1-\hat{p}_0+\Delta)}} \quad (2.2)$$

ここで $\hat{p}_0$ は帰無仮説 $H_0(\Delta)$ (2.1) の下での p_0 の最尤推定量であり, 具体的には

$$\log L=c+y_0\log p_0+(n_0-y_0)\log(1-p_0)+y_1\log(p_0-\Delta)+(n_1-y_1)\log(1-p_0+\Delta) \quad (2.3)$$

を最大にする値として求められる. いま検定は片側検定であるからサンプルから計算した U 値 (2.2) が, 標準正規分布の上側 α 点 K_α を越えるとき危険率 α で「治験薬は対照薬より有効率で Δ 以上劣ることはない」と結論できる. 通常のように $\alpha=0.05$ とすると $K_{0.05}\fallingdotseq 1.645$ である.

Table 3 経口用抗生剤二重盲検比較試験成績例

治験番号	治験薬	対照薬	対象疾患	治験薬 有効率	対照薬 有効率	検定
1	酢酸ミデカマイシン(1)	MDM	肺炎	88/101 87.1%	91/98 92.9%	N.S.
2	酢酸ミデカマイシン(2)	JM	扁桃炎	67/83 80.7%	70/80 87.5%	N.S.
3	酢酸ミデカマイシン(3)	MDM	急性中耳炎	27/44 61.4%	28/41 68.3%	N.S.
4	酢酸ミデカマイシン(3)	MDM	慢性中耳炎	27/63 42.9%	17/54 31.5%	N.S.
5	セファドロキシル(4)	CEX	急性扁桃炎	52/60 86.7%	48/57 84.2%	N.S.
6	セファドロキシル(4)	CEX	急性気管支炎	17/30 56.7%	20/32 62.5%	N.S.
7	セフロキサジン(5)	CEX	呼吸器感染症	85/131 64.9%	97/136 71.3%	N.S.
8	バカンピシリン(6)	AMPC	肺炎	39/57 68.4%	45/59 76.3%	N.S.
9	セファトリジン(7)	CEX	複雑性尿路感染症 1g投与	49/70 70.0%		N.S.
			複雑性尿路感染症 2g投与	47/66 71.2%	48/61 78.7%	
10	SF-837(8)	キタサミシン	歯性感染症	51/66 77.3%	58/68 85.3%	N.S.

ところで2項分布比較の場合は (2.2) 式のように直接標本平均を比較するよりは，逆正弦変換を用いる方が広い p_0 の範囲で正規近似がよいことが知られている．その原理は，y が2項分布 $\binom{n}{y}p^y(1-p)^{n-y}$ に従っているとき，$\sin^{-1}\sqrt{y/n}$ が漸近的に正規分布 $N(\sin^{-1}\sqrt{p}, 1/(4n))$ に従うとして確率計算を行えるというものである．この原理から導かれる検定統計量は次のようになる．

$$U_2=\frac{\sin^{-1}\sqrt{y_1/n_1+\Delta}-\sin^{-1}\sqrt{y_0/n_0}}{\sqrt{\frac{1}{4n_0}+\frac{1}{4n_1}\cdot\frac{(\hat{p}_0-\Delta)(1-\hat{p}_0+\Delta)}{\hat{p}_0(1-\hat{p}_0)}}} \qquad (2.4)$$

さて，このような検定方式を実際のデータに適用するにはあらかじめ Δ を与える必要がある．Δ を具体的に定めるには疾患の種類や出現頻度あるいはすでに有効な標準薬があるかないかなど様々な角度からの議論が必要である．以下では普通の状況，すなわちあまり特殊な疾患ではなく，ある程度効く標準薬がある場合を想定する．そのような状況で Δ を0.1を越えてとることは考えにくいので，以下では $\Delta=0.1$ および0.05の場合を検討する．

検定の有意水準は片側0.05と定める．すなわち (2.2) あるいは (2.4) 式の参照値として $K_{0.05}=1.645$ を採用する．

ここでこの方式がどのように働くか実例であたってみることにする．Table 3 は比較的最近，検定の結果対照薬に対し有意差が無いということで認められた経口用抗生剤の二重盲検比較試験の例である．この表に関する引用文献(1)～(8)は本文中の文献の後に一括して掲載する．なお，これらの例のうち治験番号10以外の原表はすべて Excellent, Good, Fair, Poor, Unknown の計数分類データで与えられている．本節ではこれを数値例として利用するため，Excellent+Good を有効と定め，Unknown は例数から削除した．したがって有効性判定規準や Unknown 例の取り扱いによっては違った結果になり得ることを断っておく．

まず治験番号1の例について考える．Table 3 から

$$\frac{y_1}{n_1}=\frac{88}{101}=0.871,\quad \frac{y_0}{n_0}=\frac{91}{98}=0.929$$

である．$\hat{p}_0$ は (2.3) 式を最大にする値として求められるが，Table 3 にある治験例のように例数 n_0, n_1 がほぼ等しく，かつ見かけ上有効率にあまり差がないような場合には単純な二つの推定量 $\tilde{p}_0$, $\tilde{p}_0$

$$\tilde{p}_0=\frac{y_0}{n_0},\quad \tilde{p}_1=\tilde{p}_0-\Delta=\frac{y_1}{n_1} \qquad (2.5)$$

を平均して

$$\hat{p}_0=\frac{1}{2}(\tilde{p}_0+\tilde{p}_0)=\frac{1}{2}\left(\frac{y_0}{n_0}+\frac{y_1}{n_1}+\Delta\right) \qquad (2.6)$$

としても十分良い近似値が得られる．以下では単純な推定量 $\tilde{p}_0$ を用いた場合と $\hat{p}_0$ を用いた場合の両方について統計量 U_1 (2.2)，U_2 (2.4) を計算するが，とくに U_2 はどちらの推定量を用いたかによる変化が少なく，安定した統計量であることがわかる．なお，Dunnet & Gent (1977) は U_1((2.2) 式) で

$$\hat{p}_0=\frac{y_0+y_1+n_1\Delta}{n_0+n_1} \qquad (2.7)$$

とし，連続修正を用いる方法を提案し，Gart (1971) に

よる条件付推測（y_0+y_1 を与えた条件付分布は一般超幾何分布になる）との比較を行っている．2群でのサンプル・サイズがほぼ等しいとき (2.6) 式と (2.7) 式はほとんど同等である．

この例の場合 $\tilde{p}_0=0.929$ に対し，

$$\hat{p}_0=\frac{1}{2}(0.929+0.871+0.1)=0.950 \qquad (2.8)$$

となる．これらから統計量 U_1 を計算すると次のようになる．

$$U_1(\tilde{p}_0)=\frac{0.871+0.1-0.929}{\sqrt{\frac{1}{98}\cdot 0.929(1-0.929)+\frac{1}{101}\cdot 0.829(1-0.829)}}=0.922$$

$$U_1(\hat{p}_0)=\frac{0.871+0.1-0.929}{\sqrt{\frac{1}{98}\cdot 0.95(1-0.95)+\frac{1}{101}\cdot 0.85(1-0.85)}}=1.01$$

次に (2.4) 式によって U_2 を計算すると次のようになる．なお，$\sin^{-1}\sqrt{0.871+0.1}=1.400$，$\sin^{-1}\sqrt{0.929}=1.301$ である．

$$U_2(\tilde{p}_0)=\frac{1.400-1.301}{\sqrt{\frac{1}{4\times 98}+\frac{1}{4\times 101}\cdot\frac{(0.929-0.1)(1-0.929+0.1)}{0.929(1-0.929)}}}=1.19$$

$$U_2(\hat{p}_0)=\frac{1.400-1.301}{\sqrt{\frac{1}{4\times 98}+\frac{1}{4\times 101}\cdot\frac{(0.95-0.1)(1-0.95+0.1)}{0.95(1-0.95)}}}=1.03$$

この結果 $\hat{p}_0$ (2.6) を用いた場合に U_1 と U_2 との非常によい一致が得られているが，これは以下に示す例でも共通に見られる現象である．検定の参照値は $K_{0.05}=1.645$ だから得られた U 値は有意ではない．この例では見かけ上治験薬が有効率で約 0.06 劣っており，ばらつきを考慮すると 0.1 劣ることがないという十分な証拠が得られていないことを意味する．やや乱暴な議論だが治験を拡大したときに現在の有効率比が保たれると仮定すると，現在の約 $(1.645/1.03)^2=2.55$ 倍の例数があれば有意差を検出できることになる．

次に差 $\Delta=0.05$ としたときの計算を行う．今度は (2.6) 式の $\hat{p}_0$,

$$\hat{p}_0=\frac{1}{2}(0.929+0.871+0.05)=0.925$$

を用いた場合についてのみ示す．まず (2.2) 式から

$$U_1(\hat{p}_0,\ \Delta=0.05)=\frac{0.871+0.1-0.929}{\sqrt{\frac{1}{98}\cdot 0.925(1-0.925)+\frac{1}{101}\cdot 0.825(1-0.825)}}=-0.19$$

が得られる．次に $\sin^{-1}\sqrt{0.871+0.05}=1.286$ であるから (2.4) 式より

$$U_2(\hat{p}_0,\ \Delta=0.05)=\frac{1.286-1.301}{\sqrt{\frac{1}{4\times 98}+\frac{1}{4\times 101}\cdot\frac{(0.925-0.1)(1-0.925+0.1)}{0.925(1-0.925)}}}=-0.19$$

が得られる．ふたたび U_1, U_2 のよい一致が見られる．今度は現在の比率差（約 0.06）が容認する $\Delta=0.05$ より大きいので，このままの比率が保たれるといくらサンプル・サイズを拡大しても同等性はいえないことになる．さて次に治験薬が見かけ上わずかながら優っているセファドロキシル対 CEX（急性扁桃炎）の例を考える．まず $\Delta=0.1$ とする．

Table 3（治験番号 5）より，

$$\frac{y_1}{n_1}=\frac{52}{60}=0.867,\quad \frac{y_0}{n_0}=\frac{48}{57}=0.842(=\tilde{p}_0)$$

である．$\hat{p}_0$ は (2.6) 式より近似的に

$$\hat{p}_0=\frac{1}{2}(0.867+0.842+0.1)=0.904 \qquad (2.9)$$

と得られる．以上から

$$U_1(\tilde{p}_0)=\frac{0.867+0.1-0.842}{\sqrt{\frac{1}{57}\cdot 0.842(1-0.842)+\frac{1}{60}\cdot 0.742(1-0.742)}}=1.68$$

$$U_1(\hat{p}_0)=\frac{0.867+0.1-0.842}{\sqrt{\frac{1}{57}\cdot 0.904(1-0.904)+\frac{1}{60}\cdot 0.804(1-0.804)}}=1.94$$

が得られる．$U_1(\tilde{p}_0)$ と $U_1(\hat{p}_0)$ の間に若干の食い違いが見られる．次に，$\sin^{-1}\sqrt{0.867+0.1}=1.388$，$\sin^{-1}\sqrt{0.842}=1.162$ だから (2.4) 式より

$$U_2(\tilde{p}_0)=\frac{1.388-1.162}{\sqrt{\frac{1}{4\times 57}+\frac{1}{4\times 60}\cdot\frac{(0.842-0.1)(1-0.842+0.1)}{0.842(1-0.842)}}}=2.22$$

$$U_2(\hat{p}_0)=\frac{1.388-1.162}{\sqrt{\frac{1}{4\times 57}+\frac{1}{4\times 60}\cdot\frac{(0.904-0.1)(1-0.904+0.1)}{0.904(1-0.904)}}}=2.07$$

Table 4 同等性検定の統計量（各治験で上段が対照薬，下段が治験薬）

治験番号	薬剤名	対象疾患	サンプルサイズ	標本有効率	標本有効率差（治験−対照）	H_0 ($\Delta=0.1$) の下で評価した標準偏差	$U_1(\tilde{p}_0)$	$U_1(\hat{p}_0)$	$U_1(\hat{p}_0, \Delta=0.05)$	$U_2(\tilde{p}_0)$	$U_2(\hat{p}_0)$	$U_2(\hat{p}_0, \Delta=0.05)$	必要なサンプル・サイズ比	
													$\Delta=0.1$	$\Delta=0.05$
1	MDM 酢酸ミデカマイシン	肺炎	98 101	0.929 0.871	−0.058	0.042	0.92	1.01	−0.19	1.19	1.03	−0.19	2.6倍	——
2	JM 酢酸ミデカマイシン	扁桃炎	80 83	0.875 0.807	−0.068	0.057	0.54	0.57	−0.32	0.58	0.57	−0.32	8.3倍	——
3	MDM 酢酸ミデカマイシン	急性中耳炎	41 44	0.683 0.614	−0.069	0.103	0.30	0.30	−0.18	0.30	0.30	−0.18	30倍	——
4	MDM 酢酸ミデカマイシン	慢性中耳炎	54 63	0.315 0.429	0.114	0.089	2.62**	2.40**	1.83*	2.49**	2.42**	1.84*	0.46倍	0.80倍
5	CEX セファドロキシル	急性扁桃炎	57 60	0.842 0.867	0.025	0.064	1.68*	1.94*	1.16	2.22*	2.07*	1.17	0.63倍	2.0倍
6	CEX セファドロキシル	急性気管支炎	32 30	0.625 0.567	−0.058	0.124	0.33	0.34	−0.07	0.34	0.34	−0.07	23倍	——
7	CEX セフロキサジン	呼吸器感染症	136 131	0.713 0.649	−0.064	0.057	0.62	0.63	−0.25	0.63	0.63	−0.25	6.8倍	——
8	AMPC バカンピシリン	肺炎	59 57	0.763 0.684	−0.079	0.083	0.26	0.26	−0.34	0.26	0.26	−0.34	40倍	——
9	CEX セファトリジン (2g投与)	複雑性尿路感染症	61 66	0.787 0.712	−0.075	0.076	0.33	0.33	−0.33	0.33	0.33	−0.32	25倍	——
10	キタサミシン SF-837	歯性感染症	68 66	0.853 0.773	−0.080	0.067	0.26	0.30	−0.42	0.26	0.30	−0.42	30倍	——

が得られる．$U_2(\hat{p}_0)$ と $U_2(\hat{p}_0)$ の食い違いはそう大きくないし，$U_1(\hat{p}_0)$ と $U_2(\hat{p}_0)$ はよく一致している．いずれにせよ参照値 $K_{0.05}$ と比較して有意差が示される．

次にこの例で $\Delta=0.05$ としてみよう．$\hat{p}_0$ は

$$\hat{p}_0=\frac{1}{2}(0.867+0.842+0.05)=0.879$$

と変更される．(2.2) 式と (2.4) 式で $\Delta=0.05$ とおいて前と同じように計算すると，

$U_1(\hat{p}_0)=1.16$

$U_2(\hat{p}_0)=1.17$

が得られる．$\Delta=0.05$ に対しては本治験は十分な保証を与えていないことがわかる．現在の有効率比が保たれるとして，有意差を示すに要するサンプル・サイズは，$(1.645/1.17)^2=1.98$ から，現在の約2倍であることがわかる．

さてこの他の例については計算結果を一括して Table 4 に掲げる．Table 4 には $\Delta=0.1$ に対する $U_i(\hat{p}_0)$，$U_i(\hat{p}_0)$，および $U_i(\hat{p}_0,\ \Delta=0.05)$，$i=1.2$，のほか，参考として標本有効率とその差，差に対する標準偏差を $H_0(\Delta)$: $p_1=p_0-0.1$ の下で評価した値を示す．さらに最後の列には標本有効率がこのまま保たれたと仮定したときに有意差を検出するのに必要なサンプル・サイズ比を示してある．

Tabl 4 を概括して受ける印象として $\hat{p}_0$ と $\bar{p}_0$ を用いたときの統計値の差は小さく安定している．それはとくに U_2 において顕著である．また，$\hat{p}_0$ を用いたとき U_1 と U_2 は極めてよく一致している．このことからとりあえず統計値としては (2.4) 式，$\hat{p}_0$ としては (2.6) 式を用いればよいと思われる．ただし両群で例数が大きく異なるときは (2.3) 式の最大化によって $\hat{p}_0$ を求めることが望ましい．

容認できる差 Δ の選択は難しい問題である．治験番号4,5の結果を見ると対照薬と真に同等かそれ以上のものを拾うのであれば $\Delta=0.05$ でよいように思えるが，他の側面で確かに優れていて有効率ではやや劣るものも拾うという主旨であれば $\Delta=0.10$ でもよいと思う．この辺は経験をつみかさね，また，情況に応じて決定されるべきであろう．

さらに詳細に表を見てみよう．治験番号2，3は見かけ上の有効率の差は -0.068，-0.069 と値が似通っている．しかるにサンプル・サイズは一方が他方の約2倍であり，これは統計値に約$\sqrt{2}$倍の影響をもたらす．さらに治験番号2と3では $\hat{p}_0\fallingdotseq0.89$ および $\hat{p}_1\fallingdotseq0.70$ の差がある．標本平均の標準偏差 $\sqrt{p(1-p)/n}$ は $p=\frac{1}{2}$ で最大で p が0,1に近づくにつれ小さくなる．その意味で有効率の高い治験2の方が統計値のばらつきが小さい．これをサンプル・サイズに換算すると大雑把にいって $\sqrt{0.70\times0.30/(0.89\times0.11)}=1.46$ 程度である．これと本来のサンプル・サイズの比が相乗されて統計値は $\sqrt{2}\times1.46\fallingdotseq2$ 倍くらい変ってくるはずである．これは荒い見積りであるが Table 4 とよく対応している．これが，見かけ上の差がほぼ等しいにもかかわらず，必要なサンプル・サイズの欄で約 $4(2^2)$ 倍の差異が生じていることの説明になる．治験番号1と6の比較でもまったく同じことがいえる．概して有効率が0.85を越えるようであれば多少対照薬より劣っていても，$\Delta=0.1$ を上乗せすることによって通常のサンプル・サイズで有意差が出そうである．

Table 5 $p_0=p_1$ のときに必要なサンプル・サイズ
(有意水準 $\alpha=0.05$(片側)，第2種の過誤 $\beta=0.10$)

$p_0=p_1$	0.90	0.85	0.80	0.70	0.60	0.50
$\Delta=0.1$	77	201	263	353	405	423
$\Delta=0.05$	596	862	1088	1433	1639	1708

つぎに，薬剤の有効率が真に等しいとき，この Δ を上乗せして片側有意水準 $\alpha=0.05$ の検定を行う方式によって同等以上であることを保証できる確率が0.9以上(第2種の過誤の確率 β が0.1以下)となるために必要なサンプル・サイズを求めてみる．サンプル・サイズは両薬剤共通とし，n とおく．これは (2.4) 式で $n_0=n_1=n$ とし，y_0, y_1, $\hat{p}_0$ をそれぞれ期待値で置き換えたものが $K\alpha+K_\beta$ を越えるとの要請から定められる(たとえば，広津，1983参照)．$\alpha=0.05$，$\beta=0.1$ に対しては $K_{0.05}+K_{0.1}=1.645+1.282=2.927$ である．いま，$p_0=p_1=p$ とおくと，$E(y_0)=E(y_1)=np$,

$$E(\hat{p}_0)=E\left\{\frac{1}{2}\left(\frac{y_0}{n_0}+\frac{y_1}{n_1}+\Delta\right)\right\}=\frac{1}{2}(p+p+\Delta)$$

$$=p+\frac{1}{2}\Delta \qquad (2.10)$$

となる．以上から求めるサンプル・サイズは

$$n>\frac{2.927^2\left\{1+\dfrac{(p-\frac{1}{2}\Delta)(1-p+\frac{1}{2}\Delta)}{(p+\frac{1}{2}\Delta)(1-p-\frac{1}{2}\Delta)}\right\}}{4(\sin^{-1}\sqrt{p+\Delta}-\sin^{-1}\sqrt{p})^2} \qquad (2.11)$$

で与えられる．Table 5 に $p=0.9\sim0.5$ の場合について (2.11) 式を計算した値を示す．上段が $\Delta=0.1$，下段が $\Delta=0.05$ の場合の値である．この範囲で $\Delta=0.1$ に対するサンプル・サイズは合理的に見える．$\Delta=0.05$ に対する値はやや大きく見えるかも知れないが通常の検定 ($\Delta=0$, $K_\alpha=1.96$ に相当) では $n=\infty$ で $\beta=0.5$ であ

Table 6 必要なサンプル・サイズ
($\Delta=0.1$, $\alpha=0.05$, $\beta=0.10$)

(1) p_1 が p_0 より 0.05 優っているとき (* $p_0=0.90$ の場合のみ $\Delta=0.05$ とした)

	p_0 / p_1	0.90 / 0.95	0.85 / 0.90	0.80 / 0.85	0.70 / 0.75	0.60 / 0.65	0.50 / 0.55
n	同等性検定	60*	42	98	147	175	187
	通常の検定	568	743	985	1364	1605	1708

(2) p_1 が p_0 より 0.05 劣っているとき

	p_0 / p_1	0.95 / 0.90	0.90 / 0.85	0.85 / 0.80	0.80 / 0.75	0.70 / 0.65	0.60 / 0.55	0.50 / 0.45
n	同等性検定	231	713	963	1173	1478	1656	1691
	通常の検定	真値で少しでも劣っていると n をいくら増しても（増す程）positive な結果は得られない						

ることを考えればずっと小さな値である．

つぎに真の p_1 が p_0 に対し0.05優っているときと劣っている場合について同様の計算をしてみよう．そのために (2.11) の公式は次のように拡張される．

$$n>\frac{2.927^2\left\{1+\frac{(p_0+p_1-\Delta)(2-p_0-p_1+\Delta)}{(p_0+p_1+\Delta)(2-p_0-p_1-\Delta)}\right\}}{4(\sin^{-1}\sqrt{p_1+\Delta}-\sin^{-1}\sqrt{p_0})^2} \tag{2.12}$$

$\Delta=0.1$ とした場合の結果を Table 6 に掲げる．表中には参考として通常の検定で要求されるサンプル・サイズも示してある．

Table 6 の結果からも，対照薬の有効率が 0.9～0.5 の範囲では $\Delta=0.1$ が合理的のように見える．有効率の高いところでのサンプル・サイズは現状で十分だし，有効率の低いところでも法外な要求にはなっていない（表の下段に示した通常の検定で要求されるサンプル・サイズと比較して欲しい）．有効率の低いところでの要請が若干きつくなるのはむしろ自然であろう．対照薬の有効率が 0.9 を越えるようだと $\Delta=0.05$ ぐらいの方がよいかも知れない．逆に，標準薬の有効率が 0.5 に満たず，新薬の開発が急がれているとき，あるいは稀な症例でサンプル・サイズが容易に集められないような場合には別途考慮が必要である．

なお，ここで述べた同等性検定のサンプル・サイズ評価は Makuch & Simon (1978) や Detsky & Sackett (1985) らも行っているが，非心度の評価の仕方がここで用いた方法と若干異なっている．すなわち，Makuch & Simon は標準偏差の計算に $p_0=p_1$ を仮定しており，Detsky & Sackett は $\hat{p}_0=p_0$, $\hat{p}_1=p_1$ としているがいずれを用いてもそう極端な違いはないと思われる．

現行の，有意でなければ採択という方式だと良い治験薬が誤って却下される確率はサンプル・サイズと無関係に有意水準 α 以下である．一方審査側としては，劣っている治験薬を却下するためには有意差を示さないといけないので，法外なサンプル・サイズを要求せざるを得ない．これは真に良い薬剤を世に出すのには費用，手間の面で大きな妨げになる．これに対し，ここで述べた同等性検定は有意差が出れば採択ということなので審査側はサンプル・サイズを規定する必要がない．サンプル・サイズは受審側の裁量に委ねられるが，たとえば $\Delta=0.1$ とすると真に有効な治験薬ならサンプル・サイズは現状よりむしろ少めで済む，相対的に有効率の低い治験薬ではやや多めになるが，この場合より慎重な検討が要求されるのは当然だろう．

Δ を上乗せする方式はどちらかというと差の区間推定の考え方に近づいている．普通の有意差検定方式は対応する信頼区間が 0 を含むか含まないかの 0，1 で判定し，0 を含んでいればいくら大きな差が示唆されていても（信頼区間の幅がいくら広くても）それはあまり問題にしない（検出力はあまり考慮されていないので）．これに対し Δ 上乗せ方式はネガティブな差がある程度以内であることを積極的に示すために，信頼区間の幅が問題になる．つまり，区間推定で信頼区間の広さを問題にしている．この点については次回 2.2) 節でさらに詳しく述べることにする．

なお，本年 8 月初めに米国 FDA (Food and Drug

Administration), を訪問し, Dr. R. Temple, Dr. S. Dubey 他統計家3名と臨床試験の様々な問題点について討論した. 同等性検証もその重要議題の一つであったが, FDA では信頼区間の幅に規準を設けていることを知った. FDA 方式とここで紹介した Δ 上乗せ方式はいろいろな意味で非常に似通っているということで意見が一致したが, 直観的には Δ 上乗せ方式の方が理解しやすいのではないだろうか.

文 献

1) Armitage, P.: Two areas of controversy in the design and analysis of clinical trials. Paper read at a meeting on 'Recent Topics on Clinical trials' held in Tokyo, 8 September, as a satellite meeting to the 12th International Biometric Conference, 1984.
2) Detsky, A.S. & Sackett, D.L.: When was a 'negative' clinical trial big enough? *Arch. Intern. Med.*, 145: 707-712, 1985.
3) Dunnet, C.W. & Gent, M.: Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics*, 33: 593-602, 1977.
4) 藤田利治, 椿 広計, 佐藤倚男, 栗原雅直, 藤本聡：多数の反応項目およびサブグループに対する統計的検定の繰返し適用の問題点. 臨床評価, 12: 827-835, 1984.
5) 藤田利治, 椿 広計, 佐藤倚男：臨床試験における多重性, 多重比較を中心として. 臨床評価, 14: 301-309, 1986.
6) Gart, J.J.: The comparison of proportions: a reiew of significance tests, confidence intervals and adjustments for stratification. *Review of Internationa Statistical Institute*, 39: 148-169, 1971.
7) 広津千尋：統計的データ解析—工学, 医学, 薬学, 社会データの実例による説明—. 日本規格協会, 東京, 1983.
8) 広津千尋：薬効検定でよく用いられる統計的方法とその問題点について. 臨床評価, 12: 309-319, 1984 a.
9) 広津千尋：薬効検定でよく用いられる統計的方法とその問題点について (2). 臨床評価, 12: 589-610, 1986 b.
10) Makuchi, R. & Simon, R.: Sample size requirements for ealuating a conservative therapy. *Cancer Treat. Rep.*, 62: 1037-1040, 1978.
11) Pocock, S.J.: Cinical trias—A practicl approach—. John Wiley & Sons, New York, 1984a.
12) Pocock, S.J.: Current issues in the design and interpretation of clinical trials. Proceedings of the 12th Internationa Biometric Conference, Invited Papers: 31-39, 1984b.

Table 3 の引用文献

(1) 感染症学雑誌, 56(11): 1045-1089, 1982.
(2) 感染症学雑誌, 56(11): 982-1002, 1982.
(3) *Chemotherapy*, 31(4): 411-434, 1983.
(4) *Chemotherapy*, 29(1): 30-47, 1981.
(5) *Chemotherapy*, 28(7): 918-963, 1980.
(6) *Chemotherapy*, 27(5): 725-759, 1979.
(7) *Chemotherapy*, 26(1): 1-9, 1978.
(8) 日本口腔外科学会誌, 18(3): 395-404, 1972.

* * *