Clin. Eval. 14 : 707-722, 1986. Dec.

総　説

臨床試験における統計的諸問題 (2)

——統計的検定の多重性について——

広 津 千 尋*

Some Statistical Problems in Clinical Trials (2)

——Multiplicity of Statistical Tests——

Key words: Cumulative chi-sequared test; F test; Dunnet's test; Kruscal & Wallis' test; Maximum chi-squared test; Pearson's chi-sequared test; Tukey's test; Wilcoxon, Mann & Whitney's test

Chihiro Hirotsu: Faculty of Engineering, University of Tokyo.

The problems of applying two or more statistical tests to a set of data is discussed. It occurs:

(1) in analyzing ordered categorical data, where there are some choices of statistical tests according to the treatment of the ordered categories, and

(2) in comparing two or more drugs simultaneously.

For case (1) it is pointed out that the most popular Wilcoxon-Mann-Whitney test may be too sensitive against the types of alternatives, and the cumulative chi-squared test, which keeps relatively high power against the wide range of ordered alternatives, is recommended in circumstances where less is known about the type of treatment effects.

For case (2) an appropriately specified multiple compaisons procedure is recommended in place of the well known omnibus tests such as the F or the Kruscal-Wallis.

In any case it is most desirable to specify the hypotheses and the testing procedures in the protocol in advance to obtaining data.

* 東京大学工学部計数工学科

2. ネガティブ・トライアルの考え方に基づく同等性検証

2.2) Δ 上乗せ検定方式と通常の帰無仮説検定方式との関係

本法が通常の帰無仮説検定方式とどのような関係にあるかを具体的に $\Delta=0.1$ として調べて見よう．

例としてまず治験番号1を考える．帰無仮説

$$H_0: p_1=p_0 \qquad (2.13)$$

の下で，共通の有効率推定値は

$$\hat{p}_0=\hat{p}_1=\frac{91+88}{98+101}=0.8995$$

である．これから標本有効率差に対する分散は

$$\mathrm{var}\left(\frac{y_1}{n_1}-\frac{y_0}{n_0}\right)=\left(\frac{1}{101}+\frac{1}{98}\right)0.8995(1-0.8995)$$
$$=0.00182$$

と推定できる．これから標準偏差の推定値は

$$\mathrm{sd}\left(\frac{y_1}{n_1}-\frac{y_0}{n_0}\right)=0.0427$$

と得られる．したがって有効率差に対する有意水準0.05の両側検定の棄却限界は

$$\pm 0.0427\times 1.96=\pm 0.0836 \qquad (2.14)$$

となる．棄却域は Fig. 1 斜線部のようになる．有効率差の実現値 (88/101−91/98=−0.0573) は採択域の中にある．

一方，Δ 上乗せ検定方式の帰無仮説

$$H_0(\Delta=0.1): p_1=p_0-0.1 \qquad (2.15)$$

の下での標本有効率差の分散は (2.2) 式の分母にある計算式によって

$$\mathrm{var}\left(\frac{y_1}{n_0}-\frac{y_0}{n_0}\right)=\left\{\frac{1}{98}0.950(1-0.950)\right.$$
$$\left.+\frac{1}{101}(0.850)(1-0.850)\right\}$$
$$=0.00175$$

と推定できる．ただし，$\hat{p}_0=0.950$ は(2.8)式で得られている．これから標準偏差の推定値は $\sqrt{0.00175}=0.0418$ と得られる (Table 4 では四捨五入した値が示してある)．したがって $H_0(\Delta=0.1)$ の，片側対立仮説 $H: p_1>p_0-0.1$ に対する，有意水準 0.05 の検定の棄却限界は

$$0.0418\times 1.645-0.1=0.0688-0.1=-0.0312$$

となる．この棄却域は Fig. 1 に横線で示した．実現値は再び採択域の中にある．すなわち，この例では実現値がいくら H_0 (2.13) の採択域内にあるとはいえ，その採択域が広過ぎて有効率差が −0.1 以内 (治験薬の有効率が対照薬より 0.1 以上劣ることがない) という積極的な保証が得られないことになる．このような場合，治験薬が標準薬に対し同等以上であるとはいえないだろう．

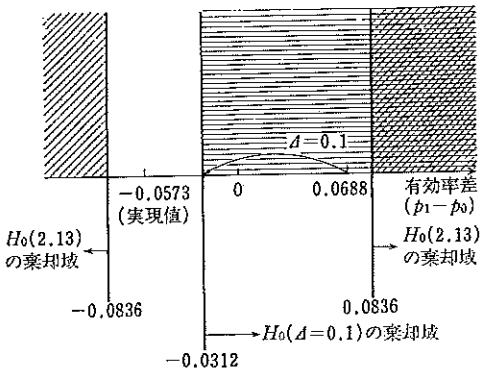

Fig. 1 治験番号1の場合の棄却域と実現値

次に治験番号5について考えよう．まず，帰無仮説 H_0 (2.13)の下で，共通の有効率推定値は

$$\hat{p}_0=\hat{p}_1=\frac{48+52}{57+60}=0.8547$$

である．これから標本有効率差の分散は

$$\mathrm{var}\left(\frac{y_0}{n_1}-\frac{y_0}{n_0}\right)=\left(\frac{1}{60}+\frac{1}{57}\right)0.8547(1-0.8547)$$
$$=0.00425$$

と推定できる．標準偏差の推定値は $\sqrt{0.00425}=0.0652$ となる．これから標本有効率差に対する有意水準 0.05 の両側検定の棄却限界は

$$\pm 0.0652\times 1.96=\pm 0.128$$

と得られる．この棄却域は Fig. 2 に斜線で示す．これは Fig. 1 よりさらに広い棄却域で，有効率差の実現値 (52/60−48/57=0.025) が採択域に入ったからといって，それだけでは p_1-p_0 に対しさしたる保証を与えることはできない．次に $H_0(\Delta=0.1)$ (2.15) の下での標本有効率差の分散は

$$\mathrm{var}\left(\frac{y_1}{n_1}-\frac{y_0}{n_0}\right)=\left\{\frac{1}{57}0.904(1-0.904)\right.$$
$$\left.+\frac{1}{60}0.804(1-0.804)\right\}=0.00415$$

と得られる．ただし，$\hat{p}_0=0.904$ は (2.9) 式で得られたものである．これから標準偏差の推定値は $\sqrt{0.00415}=0.0644$ となる．したがって対立仮説 $H_1: p_1>p_0-0.1$ に対する有意水準 0.05 の棄却限界が

$$0.0644\times 1.645-0.1=0.1060-0.1=0.0060$$

と得られる．この棄却域は Fig. 2 に横線で示す．

今度は実現値が $H_1: p_1-p_0>-0.1$ に対する棄却域の中に入っているから，治験薬が対照薬より 0.1 以上劣ることはないという積極的な保証が得られたことになる．

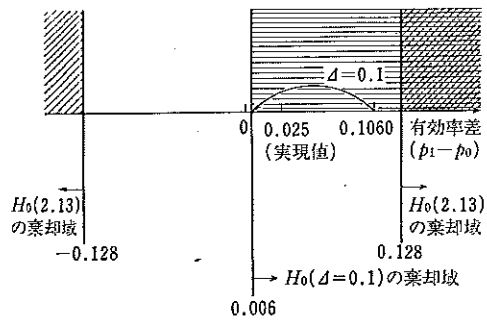

Fig. 2 治験番号5の場合の棄却域と実現値

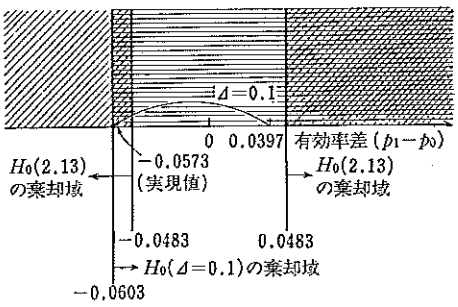

Fig. 3 Table 7 のデータに対する棄却域と実現値

Tzble 7 治験番号1を3倍に拡大したもの

	サンプル サイズ	有効症例	無効症例
対照薬	249	273	21
治験薬	303	264	39

最後に架空の例として治験番号1のデータがそのままの比率でサンプル・サイズが3倍に拡大されたものを考えよう (Table 7参照).

この場合標本有効率差はすでに求めたように −0.0573である. 標準偏差は前の $1/\sqrt{3}$ になるから, (2.14) 式の棄却限界は

$$\pm 0.0427\times\frac{1}{\sqrt{3}}\times 1.96=\pm 0.0483$$

と変更される. この棄却域は Fig. 3 に斜線で示す. 同様に $H_0(\Delta=0.1)$ (2.15) の下での標準偏差も

$$0.0418\times\frac{1}{\sqrt{3}}=0.0241$$

と変更される. したがって, 片側対立仮説 $H_1: p_1>p_0-0.1$ に対する有意水準0.05の棄却限界は

$$0.0241\times 1.645-0.1=0.0397-0.1=-0.0603$$

となる. この棄却域は Fig. 3 に横線で示した. サンプル・サイズの拡大によって検定がシャープになり両方の検定とも有意になることが分る (Fig. 1 と見比べて欲しい). 2.1) 節の検定で治験薬の対照薬に対する同等性がいえたのに, 通常の検定で有意に負けてしまうというのが一見奇妙に思えるかも知れない. しかし, 2.1) 節の検定が有意ということは治験薬が対照薬に有効率で0.1以上負けることがないことを危険率0.05で保証するものであるから, 下駄無しでは負けることがまた危険率0.05でいえたとしてもこれらは決して矛盾していない. むしろサンプル・サイズが十分大きいとこのようなことはときどき生じると考えた方がよい. このようなとき治験薬を認めるか認めないかはあらかじめ決めておけばよい. ただし, 他の諸特性を考慮して認可の決定をするためには, 事前に相当の議論が費やされるべきである. また, 有効率差の推定値も公表する (単に公表というよりは周知させる) べきである. きわめて精度のよいネガティブな推定値と, それにもかかわらず認可された理由とが明示されていれば実際上あまり問題は生じないだろう.

2.3) 2薬剤のサンプル・サイズがほぼ等しいときの近似について

2薬剤のサンプル・サイズがほぼ等しいとき, Δ 上乗せ片側5%検定方式で棄却されることと, $H_0: p_1=p_0$ (2.13) の下での標準偏差推定値に基づいた p_1-p_0 に対する信頼率90% の信頼区間が $-\Delta$ より右寄りであることとはほぼ等しい. このことを以下に示そう.

まず, Δ 上乗せ片側5%検定方式で $V_1(\hat{p}_0)$ に基づくものは, その構成からいって,

$$H_0(\Delta): p_1=p_0-\Delta \tag{2.1}$$

の下での標準偏差推定値

$$\sqrt{V_\Delta}=\sqrt{\frac{1}{n_0}\hat{p}_0(1-\hat{p}_0)+\frac{1}{n_1}(\hat{p}_0-\Delta)(1-\hat{p}_0+\Delta)} \tag{2.16}$$

に基づいた p_1-p_0 に対する信頼率90% の信頼区間が $-\Delta$ より右側であることと同値である. なぜなら (2.2) 式の U_1 が $K_{0.05}=1.645$ より大きいことと,

$$-\Delta<\frac{y_1}{n_1}-\frac{y_0}{n_0}-1.645\sqrt{V_\Delta}$$

が同値だからである. ところで, $\Delta=0.1$ の場合に

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

を用いた場合の $\sqrt{V_\Delta}$ の例が前稿 (広津, 1986) の Table 4 第7列に与えられているが, これは Table 8 に示すように, $p_1=p_0(=p)$ の下での推定値

$$\hat{p}=\frac{y_0+y_1}{n_0+n_1} \tag{2.17}$$

Table 8 標本有効率差 $(y_1/n_1-y_0/n_0)$ の二通りの標準偏差推定値の比較

治験番号	$\sqrt{V_\Delta}$ ((2.16) 式，ただし $\Delta=0.1$，$\hat{p}_0$ は (2.6) 式による	$\sqrt{V}$ (2.18) 式
1	0.042	0.043
2	0.057	0.057
3	0.103	0.104
4	0.089	0.090
5	0.064	0.065
6	0.124	0.125
7	0.057	0.057
8	0.083	0.083
9	0.076	0.077
10	0.067	0.067

に基づいた $(y_1/n_1-y_0/n_0)$ の標準偏差の推定値

$$\sqrt{V}=\sqrt{\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\hat{p}(1-\hat{p})} \qquad (2.18)$$

と大変似通っている．

n_0 と n_1 がほぼ等しいとき，このような結果が得られることは一般的に次のようにして示せる．

まず $n_0\fallingdotseq n_1$ のとき，(2.6) 式と Dunnet & Gent (1977) の推定式

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

はほぼ同等である．そこで (2.16) 式の $\hat{p}_0$ として (2.7) 式を用いると，次式が得られる．

$$V_\Delta=\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\cdot\frac{y_0+y_1}{n_0+n_1}\left(1-\frac{y_0+y_1}{n_0+n_1}\right)+\Delta\cdot\frac{n_1-n_0}{n_0n_1}\left(1-2\frac{y_0+y_1}{n_0+n_1}\right)-\Delta^2\frac{n_0^2-n_0n_1+n_1^2}{n_0n_1(n_0+n_1)} \qquad (2.19)$$

(2.19) 式の第一項は (2.18) 式の V にほかならず，第二項以下が差を表わすが，$\Delta=0.1$ で，n_0, n_1 が 50 程度だとこの差は無視できるほどに小さい．この差を V に対する相対誤差の形で表わすと，たとえば治験番号 1 の場合で -0.041，治験番号 5 の場合で -0.035 となる．標準偏差に対する相対誤差は，これらの約 1/2 でそれぞれ -0.021，-0.018 となる．

もっと一般に n_0, n_1 がほぼしいとき，それらを共通に n と表わし，差 n_1-n_0 を $\delta(n)$ と表わして標準偏差の相対誤差を評価する式を導くと次のようになる．ただし，$\hat{p}$ は (2.17) 式に与えられたものである．

$$\frac{\sqrt{V_\Delta}-\sqrt{V}}{\sqrt{V}}=-\frac{1}{2\hat{p}(1-\hat{p})}\times\left\{(2\hat{p}-1)\frac{n_1-n_0}{n_0+n_1}\Delta+\frac{n_0^2-n_0n_1+n_1^2}{(n_0+n_1)^2}\Delta^2\right\}$$
$$\fallingdotseq-\frac{\Delta}{8\hat{p}(1-\hat{p})}\left\{2(2\hat{p}-1)\frac{\delta(n)}{n}+\Delta\right\} \qquad (2.20)$$

本稿の例でとりあげたような場合 (Table 3 参照) では，すでにいくつか例示したように，(2.20) 式の値はほぼ -0.02 程度である．つまり，本稿の例のように治験薬，対照薬の例数がほぼ等しいときには，有効率差 p_1-p_0 に対する信頼区間を構成するのに，$H_0: p_1=p_0$ (2.13) を仮定するのと $H_0(\Delta=0.1)$ ((2.1) 式) を仮定するのでは区間幅の相対誤差は約 2 %程度しかない．したがって Δ 上乗せ片側 5 % 検定方式による有意性判定は，実は $H_0: p_1=p_0$ の下での標準偏差推定値から構成した p_1-p_0 の信頼率90%の信頼区間が $-\Delta$ より右寄りであることの判定とほぼ対応する．かくして，Δ 上乗せ方式と，信頼区間の幅を規制する FDA 方式との対応が明らかになった．ただし，これはあくまで $n_0\fallingdotseq n_1$ の下で示されることであるから一般の場合には必ずしも成り立たない．一般の場合に理論的に筋が通るのは尤度 (2.3) に対する最尤推定量として p_0^* を求め，それに基づく検定統計量 $U_2(p_0^*)$ を用いた Δ 上乗せ検定方式であろう．

2.4) 許容差 Δ について

前稿 (広津，1986) で，種々検討した結果，特殊な薬剤ではなく，標準薬の有効率 p_0 が 0.5～0.9 の範囲では，一応 $\Delta=0.1$ ぐらいが妥当であろうと述べた．これに対し，$\Delta=0.1$ は $p_0=0.5$ に対しては20%に当り妥当だが，他の p_0 に対しては比例的に Δ を増すのが自然ではないかとの意見を頂戴した．その考え方は，江島・他 (1982) が扱っているような，正規分布の平均に対する場合には合理的だが，有効率のように $p=1$ のところに壁があり，

その近くで猛烈な圧縮が起こって壁に近づけないような場合には話が違ってくる．たとえば $p=0.5$ と 0.9 では標準偏差で $\sqrt{0.5(1-0.5)}/\sqrt{0.9(1-0.9)}\fallingdotseq 1.7$ 倍の違いがある．したがって有意差検定の検出力からいえば $p=0.9$ での差 Δ は $p=0.5$ での同じ差 Δ の 1.7倍に相当する．すなわち，両者で同じ Δ を用いることが，$p=0.9$ の方でほぼ比例的 $(0.9/0.5=1.8)$ に大きな Δ を想定することに対応している．

2.5) 江島・他 (1982) の同等性検定について

江島・他 (1982) が生物学的同等性の試験方法についての解説の中で統計解析について言及している．そこでは二つの正規分布の平均の同等性（等分散は仮定されている）を扱っており，相対差 $d=|\mu_0-\mu_1|/\mu_0=0.2$ に対して，検出力 0.8 以上を要請することが述べられている．ただし，μ_0, μ_1 はそれぞれ対照薬，治験薬に対する特性値の母平均である．他方，たとえ検定の結果有意差がなくても，相対差が 0.2 を越えると判断されるようなら同等とは判断しないこと，およびその反対に検定で有意差が検出されても相対差が 0.2 以内におさまるようなら，その差は問題としないことを述べている．つまり推定される相対差が 0.2 を越えるか否かを有意差検定に優先する概念として述べている．

ただしこれは相対差の単純な推定値 $\hat{d}=|\bar{y}_0-\bar{y}_1|/\bar{y}_0$ ($\bar{y}_0$, $\bar{y}_1$ は両薬剤の標本平均) が 0.2 を越えるか否かで判定することを意味するものではない．なぜならもし，$\hat{d}=0.2$ となったらな，それはやや乱暴にいえば，d が 0.2 を越えることと越えないことがほぼ等しい確率で生じることを意味するからである．いいかえると，真の相対差 d が 0.2 を越えないことをある確率で保証するには，その確率に対応する信頼率で構成した d の信頼区間が ± 0.2 の内側におさまることを要請しなければならない．ただし，d の信頼区間を $\hat{d}=|\bar{y}_0-\bar{y}_1|/\bar{y}_0$ に基づいて構成するのは，実は $\hat{d}$ の分布が不安定で好ましくない．そこでこれに代わる近似的な方法として，$\mu_0-\mu_1$ に対する通常の信頼区間 $\bar{y}_0-\bar{y}_1\pm\hat{\sigma}t_{\alpha/2}$ が $\pm 0.2\times\bar{y}_0$ の範囲内にあるとき生物学的同等性がいえたと判断する基準が導かれる．ただし，$\hat{\sigma}$ は $H_0:\mu_1=\mu_0$ の下での標準偏差の推定値，$t_{\alpha/2}$ は t 分布の両側 100α %点である．

この結果は結局有意差検定に代わって信頼区間の幅で見ることを主張するものであり（幅を $\bar{y}_0$ に依存させる点に若干問題はあるが），2.3節のアプローチと通ずるものがある．計数値を扱う場合と計量値を扱う場合では根底の統計モデルの違いによって，手法の形式が異なることはやむを得ないことであり，本稿で述べた Δ 上乗せ検定方式と，江島他 (1982) の計量値に対するアプローチとが基本的には同じことを目指していると解釈されることが示されたと思う．

2.6) まとめ

サンプル・サイズが小さいとき，通常の帰無仮説検定で有意にならないことから同等性を主張するのには無理がある．採択域が広過ぎる場合に，有意にならないことが決して真の有効率差が 0 に近いことを意味するものではないからである．一方，ここで紹介した Δ 上乗せ方式で有意差が示されれば治験薬が対照薬に対し，有効率で Δ 以上劣ることのないことが危険率 α で主張できる．

この方式を機械的に適用すると Δ だけ劣った治験薬がどんどん認可されるような印象を受けるかも知れないが，実際には通常のサンプル・サイズで，$\Delta=0.1$ とすると，本文中に述べたように平均出検品質の低下はごくわずかである．他方サンプル・サイズを非常に大きくすると，機械的には有効率でほぼ Δ 劣った薬もパスするが，そのときは非常に精度のよいネガティブな推定値が得られているわけだから，よほどの理由がないかぎり市場でどんどん使われるということはないだろう．いずれにせよこの方式が，治験・対照両薬剤の差をきちんと評価する方向を指向することは確かである．

ここで述べた同等性検証は，両薬剤のサンプル・サイズが揃っているときは，信頼率 $1-2\alpha$ の信頼区間が $-\Delta$ を含むか含まないかを基準として判定することとほぼ同等である．しかし，両群でのサンプル・サイズの相対差が 1 割を越えるようであればやはり最尤法で標準偏差を推定し，有意差検定を行った方がよい．

Δ を標準薬 0 と有効率 p_0 に相対的に決めるという考え方もあり得るが，p_0 によるばらつきの違いを考慮するとむしろ，$p_0=0.5\sim 0.9$ の範囲で一定の Δ を想定する方が合理的に思える．これは江島他(1982)が扱っている正規分布の平均の差の場合と本質的に異なる点である．

3. 臨床試験解析における検定の多重性について

3.1) 2剤比較で行われるいろいろな検定

2剤比較で最もポピュラーなのは t 検定と Wilcoxon-Mann-Whitney の順位和検定であろう．以下では後者を簡単のためにU検定と呼ぶ．本当は，標準的な統計の教科書でUは標準正規分布に従う変数として伝統的に用いられており，特定の検定に対する呼称としては好ましくないように思う．たとえば Lehman(1975) でも Wilcoxon 統計量にwを用いているが，ここでは臨床家の習慣に従っておく．

t 検定やU検定の基礎については前に本誌でも解説し

Table 9 製品濃度のデータ（森口（綱），1976から引用）
（上段生データ，下段順位）

	1	2	3	4	5	6	7	8	9	10
A社	9.1	8.1	9.1	9.0	7.8	9.4	8.2	9.1	8.2	9.3
	17	6.5	17	15	2.5	20	9	17	9	19
B社	8.2	8.6	7.8	7.6	8.4	8.6	8.0	8.1	8.8	8.0
	9	12.5	2.5	1	11	12.5	4.5	6.5	14	4.5

た（広津1984a，b）し，成書も多い（たとえば広津，1983）のでここではあらためて述べない．基本的にt統計量は計量データから直接計算され，U統計量はそれを大きさに従って順位づけしたデータから計算される．たとえばTable 9のデータではt統計量は生データから $t=2.35$ と得られ，U統計量は順位データから $U=2.11$ と得られる．

Table 9で同順位（タイ）は平均順位で置き換えてある．改善の程度でクラス分けした順序分割表で用いられるU検定は，極端にタイの多い場合と考えることができ，考え方は同じである．

さて，直観的にU検定は生データを順位に変換したときに情報をすてているから，t検定に対して若干効率が悪いと考えられる．たとえば相隣る値の差が大きくても小さくても順位差は1としか評価されない．反面とび抜けて大きな外れ値があってもその統計量に与える影響は緩和されるから頑健（ロバスト）である．……というのは通り一遍の説明であって実際にどちらの統計量を使うか（あるいはどちらも使わないか）についてはもう少し説明が必要である．

たとえば血圧は計量値で与えられるからt検定の方が有効と思えるかも知れない．しかしながら血圧値200 mmHgを180 mmHgに下げるのと，160 mmHgを140 mmHgに下げるのでは同じ20 mmHgの差でも，臨床上の意味は違うかも知れない．臨床的には改善の程度でクラス分けした頻度データの方が適切かも知れない．また生データをそのまま扱うとなると，血圧値200 mmHg付近でのばらつき（測定誤差，時間的変動などの）と160 mmHg付近でのばらつきは相当違うかも知れない，臨床的処置は平均値と共にばらつきの大きさをも変えているかも知れない，そもそも計量値といっても正規分布に従っているとは限らない，等々のことが問題になる．筆者の乏しい経験でも臨床データにはたとえば尿量や尿中排泄率など正規分布よりは裾が重く（ばらつきが大きく）かつ歪んだ分布に従うものが多いようである．

そのようなとき正規分布の仮定の下で導かれた棄却限界値はミスリーディングだし，t検定の効率が高いのはあくまで仮定した正規分布が正しいときだから実際に用いるときには正規化変換などの工夫が必要である．しかるに個々の場合にどのような正規化変換を用いるかはそれ程自明ではない．一方U検定はノンパラメトリック検定だから背後の分布によらず有意水準は保証される．しかしU検定が検出力の意味で効率が高いのは，分布が対称でかつ処理効果が平均値を変える場合である．たとえば正規分布の平均値の差に関して，U検定はt検定に対し0.95の効率を持つ（同じ検出力を得るのに1/0.95＝1.05倍のサンプル・サイズがあればよい）ことはよく知られているが，ばらつきの変化，あるいは歪んだ分布での位置パラメータ変化に関しては決して効率は高くない．逆に，t検定は漸近的にはノンパラメトリックな並べ換え検定と同等なので，nがそれほど小さくなければ十分ロバストである．実は概していうとt検定とU検定の特性は似通っている．たとえば先の例題でも $t=2.35$ は自由度18のt分布の上側約1.5%点だし，$U=2.11$ は標準正規分布の上側1.7%程度であってきわめてよく一致している．

改善度のように最初から先天的な順序に従って，カテゴライズされた頻度データでは，従来U検定か，有効以上とそれ以下の2分類にまとめなおした2×2分割表の χ^2 検定，あるいは順序を無視した一様性の検定が行われてきた．

Moses, Emerson & Hosseini (1984) によれば，New England Journal of Medicine Vol. 306(1982) の全168論文のうち32論文で47例の順序カテゴリカル・データが解析されており，意外なことに，一様性のカイ二乗検定が典型的に用いられているそうである．順序カテゴリカル・データ（順序分割表）に基づいて2剤の優劣（単なる異同ではなく）を比較するのに一様性の χ^2 検定を用いるのが無意味なことは広津（1984b）で詳しく述べた通りである．

さて，標準薬と治験薬の各改善度カテゴリの生起確率をTable 10のように表わそう．

Table 10 改善度カテゴリの生起確率

改善度	1 (Terrible)	2	……	b (Excellent)
対照薬	p_{01}	p_{02}	……	p_{0b}
治験薬	p_{11}	p_{12}	……	p_{1b}

このとき，もし，

$$\frac{p_{11}}{p_{01}} \leq \frac{p_{12}}{p_{02}} \leq \cdots\cdots \leq \frac{p_{1b}}{p_{0b}} \qquad (3.1)$$

が成り立っていれば治験薬は対照薬に対し相対的に好ましいカテゴリの割合いが高いのでより優れているといえるだろう．反対にもし

$$\frac{p_{11}}{p_{01}} \geq \frac{p_{12}}{p_{02}} \geq \cdots\cdots \geq \frac{p_{1b}}{p_{0b}} \qquad (3.2)$$

が成り立てば対照薬の方が優れているといえるだろう．ところで先に述べたようにU検定は背景に正規分布のような対称な連続分布が想定され，薬剤の効果はその平均値の変化ととらえられるような場合に検出力の高い検定である．この場合観測値が順序カテゴリカル・データとして得られると，p_{1j}/p_{0j} は j に関して厳密に単調になる．たとえば典形的な例として，対照薬の特性値が $N(0, 1^2)$ に従い，それが各カテゴリの生起する割合が等確率であるような区間データとして観測される場合を考える．このとき，(1) 4カテゴリで治験薬の特性値が $N(0.8, 1^2)$ に従う場合と，(2) 6カテゴリで治験薬の特性値が $N(0.5, 1^2)$ に従う場合の p_{0j}, p_{1j} は Table 11 のようになる．Fig. 4 はこれを図示したものである．

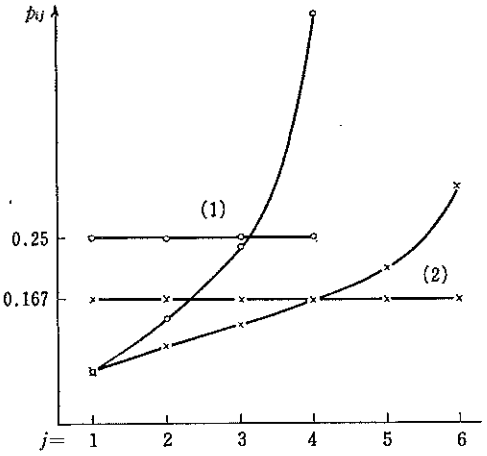

Fig. 4 Table 11 の Pij プロット

一方，歪んだ連続分布の典型として指数分布を考えて見よう．対照薬の特性値が $\exp(-u)$ 従い，治験薬ではそれが $\exp\{-(u-\mu)\}$ に従うとして先と同様に p_{ij} を構成すると Table 12 のようになる．ただし (1) では $\mu=0.2$，(2) では $\mu=0.15$ としている．

Table 11 正規分布の平均 shift で得られるセル確率 (Pij) の例

カテゴリ	1	2	3	4		
(1) 対照薬 $N(0, 1^2)$	0.250	0.250	0.250	0.250		
治験薬 $N(0.8, 1^2)$	0.070	0.142	0.238	0.550		
p_{1j}/p_{0j}	0.280	0.568	0.952	2.20		
カテゴリ	1	2	3	4	5	6
(2) 対照薬 $N(0, 1^2)$	0.167	0.167	0.167	0.167	0.167	0.167
治験薬 $N(0.5, 1^2)$	0.071	0.105	0.133	0.164	0.208	0.320
p_{1j}/p_{0j}	0.426	0.630	0.798	0.884	1.25	1.92

Table 12 指数分布の位置 sift で得られるセル確率 (Pij) の例

カテゴリ	1	2	3	4		
(1) 対照薬 ($\mu=0$)	0.250	0.250	0.250	0.250		
治験薬 ($\mu=0.2$)	0.084	0.350	0.350	0.350		
p_{1j}/p_{0j}	0.336	1.40	1.40	1.40		
カテゴリ	1	2	3	4	5	6
(2) 対照薬 ($\mu=0$)	0.167	0.167	0.167	0.167	0.167	0.167
治験薬 ($\mu=0.15$)	0.032	0.194	0.194	0.194	0.194	0.194
p_{1j}/p_{0j}	0.192	1.16	1.16	1.16	1.16	1.16

Table 13 U検定，χ^2 検定，χ^{*2} 検定の検出力比較

(1) 正規分布仮定

カテゴリ数	1群当りサンプル・サイズ	位置母数 μ		検定統計量		
		対照薬	治験薬	U	χ^2	χ^{*2}
4	80	0	0.8	0.90	0.76	0.90
6	160	0	0.5	0.89	0.66	0.89

(2) 指数分布仮定

カテゴリ数	1群当りサンプル・サイズ	位置母数 μ			検定統計量		
		対照群	治験薬1	治験薬2	U (KW)	χ^2	χ^{*2}
4	180	0	0.10	0.20	0.31	0.38	0.42
4	180	0	0.15	0.25	0.50	0.73	0.74

この場合は図にプロットするまでもなく，p_{1j}/p_{0j} が階段状に変化することが分る．式(3.1)でいえば左端を除いて不等号はすべて等号になる．このような場合U検定は後で Table 13 に示すようにむしろ順序を無視した χ^2 検定より検出力が低くなってしまう．つまり，U検定は分布形によらないノンパラメトリックな検定というものの，それは帰無仮説の下での性質であって，検出力が高い場合はかなり限られていることが分る．臨床的には正規分布 shift のように厳密に単調なパターンと指数数分布 shift で得られるような階段状のパターンのどちらが多いのであろうか．また，どちらのパターンかを事前に指定できるものだろうか．これについては諸家の意見をまたねばならないが，(3.1) あるいは (3.2) 式で表される幅広い p_{ij} の組み合わせで高い検出力を保持するものとしては，累積カイ二乗統計量 χ^{*2} が考えられる．累積カイ二乗統計量の構成については広津(1984 b)で詳しく述べたので詳細は省略する．簡単にいえばカテゴリに b−1 通りの断点を入れ，その前後でプールして，(1)×(2, …, b); (1, 2)×(3, …, b); …; (1, …, b−1)×(b) のような b−1 通りの 2×2 分割表を考え，その χ^2 を足し込んだものが χ^{*2} である．すなわち，これら b−1 通りの分割表に対する通常の χ^2 を χ_1^2, …, χ_{b-1}^2 と表すと，χ^{*2} は

$$\chi^{*2}=\chi_1^2+\cdots\chi_{b-1}^2 \qquad (3.3)$$

で与えられる．

正規分布 shift および指数分布 shift の場合にU検定，χ^2 検定，χ^{*2} 検定の検出力を比較した例（広津，1982）を Table 13 に引用する．ただし，指数分布仮定の場合は3剤の一様性比較を扱っており，たとえばU検定とあるは Kruscal-Wallis (KW) 検定のことである．Table 13 は χ^{*2} 検定が正規分布 shift の場合にU検定と comparable な検出力を持つこと，また，指数分布 shift の場合にはU検定よりはるかに大きな検出力を持ち，χ^2 検定と comparable であることを示している．この他種々行われたシミュレーションなどによっても χ^{*2} 検定が (3.1) および (3.2) 式で示される幅広い p_{ij} の範囲で高い検出力を保持することが確かめられている．検証を目的とするⅢ相試験では事前に検定手法を決めておく必要があるが，Ⅱ相までの結果によって正規分布 shift のようなパターンが十分予期できるのでなければ，χ^{*2} 検定を標準的に用いることが考えられる．

次に，臨床的に意味のある効果が事前に定義され，それ以上の効果があるかないかで分類した 2×2 分割表におけるカイ二乗検定はそれなりに合理的でとくに問題はない．またその場合は，2章で述べたような同等性検定の適用も可能である．しかし，現実によく見かけるのは，あらかじめそのような分類が与えられておらず，事後的に b−1 通りの断点で分割し，計算した χ^2 統計量の最大値を採用することである．これは (3.3) 式の χ^{*2} の最大成分を採ることにほかならず，

$$\max\chi^2=\max(\chi_1^2, \cdots, \chi_{b-1}^2)$$

と表わすことができる．ただし，max(・, …, ・) は( )内の最大成分を表わす．以下ではこれを $\max\chi^2$ 検定と呼ぶ．$\max\chi^2$ を，あたかも最初から分類の定められていた 2×2 分割表の χ^2 と見たてて，自由度1の χ^2 検定をするのが誤りであることは広津（1984 b）で詳しく述べた．事後的に b−1 通りの χ^2 値の最大を採っているのだから，それでは有意水準（第一種の過誤の確率）が増大するのは明らかである．たとえば広津（1984 b）の Table 9 によれば b=5 の場合に，$\max\chi^2$ をあたかも

自由度1の χ^2 とみなして検定すると，設定した有意水準が5%，1%の場合に，実際の有意水準はそれぞれ14%および3%を越えてしまうことが観察される．$\max\chi^2$ に対する安全側（conservative）の棄却域が Sugiura & Otake (1973) によって求められており，その近似の精度を藤田・椿 (1985) が検討している．

一方観察された $\max\chi^2$ の有意確率の上下限を与える公式 (Bonferroni の不等式) が広津 (1983) に与えられており，とくに下限の精度はかなり高い．この精度をさらに高めることも原理的には可能であるが，実は$\max\chi^2$は (3.1) や (3.2) 式で表わされるような優劣の全体的な傾向を検出するのには必ずしも適した統計量ではない．それは棄却域を図に示すとすぐ分ることだが，(3.1) あるいは(3.2) 式で表わされる領域の端での検出力を強調するあまり，領域の中央付近で棄却域がふくらんでしまうからである．つまり正規分布の平均 shift のような場合には，U検定や χ^{*2} 検定に較べてはなはだ検出力が低下してしまう（藤田・椿，1985のシミュレーション結果も参照のこと）．さらに $\max\chi^2$ のもう一つの難点はカテゴリの端で度数が極端に小さいと，そのカテゴリでの両薬剤の差を強調し過ぎる嫌いがあることである．例として Table 14 を考えよう．Table 14 から直観的に受ける印象はカテゴリ2,3の間で両薬剤のレスポンスが逆転しているということだろう．ところが Table 14 を3通りに区切って(3.3) 式の χ^2 成分を計算すると Table 15のようになり，カテゴリ3までをプールしたものとカテゴリ4との間での逆転が強調される．なぜこのようになるかというと (3) の区分表で標準偏差が極端に小さく推定されるからである．いま (2) および (3) の区分表で $\hat{p}_1-\hat{p}_0$ はそれぞれ 24/100−14/100=0.10，4/100−0/100=0.04 であり，区分 (2) の方が倍以上大きい．ところがこれらに対する標準偏差はそれぞれ

$$\sqrt{\left(\frac{1}{100}+\frac{1}{100}\right)\frac{38}{200}\left(1-\frac{38}{200}\right)}=0.0555,$$

$$\sqrt{\left(\frac{1}{100}+\frac{1}{100}\right)\frac{4}{200}\left(1-\frac{4}{200}\right)}=0.0198$$

と推定され，$\hat{p}_1-\hat{p}_0$ を標準偏差で規準化したものが1.80 (二乗が 3.25)，2.02 (二乗が 4.08) となる．つまり，(3) のような区分表ではいわば2項分布で p が1に近いときのように，標準偏差が小さく推定されるので少しの差が大きく強調されがちになる．もちろん，このように頻度の小さいカテゴリがあると，結果が不安定になること，カイ二乗近似（正規近似）がよくないことなどはよく知られていることと思うが，$\max\chi^2$ を用いるときにはとくに注意が必要である．

以上のことにより，$\max\chi^2$ 検定は優劣比較の一般的な検定としては推薦できない．とくにカテゴリの分類（多重比較）を目的とする場合か，あるいはⅡ相までの検討から効き方にくせがあり，$\max\chi^2$ が適切と判断された場合に用いるべき方法であろう．

藤田・椿 (1985) では臨床評価9(1)から13(2)までの65個の2薬剤間の比較試験結果にあらためてU検定，χ^{*2} 検定，$\max\chi^2$ 検定を適用して比較検討した結果から次のことを述べている．

① U検定でも，一様性の χ^2 検定でも有意でない場合は χ^{*2} 検定でも，$\max\chi^2$ 検定でも有意にならない．逆にU検定で高度に有意になる場合は他の2法でもやはりおおむね高度に有意になる．

② 差が出るのは（U検定で）微妙なところで有意になったり，ならなかったりした場合であるが，U検定次いで χ^{*2} 検定がよく有意差を検出しており，$\max\chi^2$ 検定はU検定で5%有意水準程度の有効差は見落し勝ちである．

Table 14　右端の頻度が小さい例

カテゴリ	1	2	3	4	計
対照薬	22	64	14	0	100
治験薬	20	56	20	4	100
計	42	120	34	4	200

Table 15　χ^2 成分の計算

	(1)		(2)		(3)		
カテゴリ	1	2, 3, 4	1, 2	3, 4	1, 2, 3	4	計
対照薬	22	78	86	14	100	0	100
治験薬	20	80	76	24	96	4	100
計	42	158	162	38	196	4	200
	⇓		⇓		⇓		
	$\chi_1^2=0.33$		$\chi_2^2=3.25$		$\chi_3^2=4.08$		

この結果は，今までに述べてきた各検定の特徴づけとも対応し興味があるが，とくに②のコメントに関しては若干注意が必要である．まず過去の治験例でU検定がやや優勢なのは，従来U検定の意味で優劣がはかられてきたことと無関係ではないだろう．すなわちⅡ相までの検討で，χ^{*2} 検定や，$\max\chi^2$ の意味で優れた薬が見落とされてきた可能性が考えられる．また，そこで論じられているのはあくまで得られたデータに対する結論が検定法によってどう変るかということであり，必ずしもどの検定結果が“真”をいいあてているかということには対応しない．検出力の理論的な検討では正規分布の平均 shift のようにU検定が優れている場合でも，χ^{*2} 検定の検出力低下はごくわずかであり，p_{1j}/p_{0j} が階段状に変化する場合でカテゴリ数が5を越えるようだと，U検定より χ^{*2} 検定の方がはるかに優れていることがわかっている．もう一つ考慮すべきことは分布の近似の問題である．これについては機会をあらためて少し詳しく述べたいと思うが，とくに両端に薄い場合は注意が必要である．

一様性のカイ二乗検定，w 検定，χ^{*2} 検定を特徴付けるものとして次のことは興味深い．すなわち，すべてのカテゴリの出現度数（対照薬，治験薬についての合計）が揃っているとき，累積カイ二乗統計量は次のように分解できる（Hirotsu, 1986）．

$$\chi^{*2}=\left\{\frac{1}{1.2}\chi^2_{(1)}+\frac{1}{2.3}\chi^2_{(1)}+\frac{1}{3.4}\times\chi^2_{(3)}+\cdots+\frac{1}{(b-1)b}\chi^2_{(b-1)}\right\}\times b$$

ただし，$\chi^2_{(1)}$，$\chi^2_{(2)}$，$\chi^2_{(3)}$，… は（チェビシェフの多項式の意味で）1次，2次，3次，…の変動を検出するための自由度1の χ^2 統計量であり，漸近的にたがいに独立である．ちなみに，他の二つは

$$U=\chi^2_{(1)}$$

$$\chi^2=\chi^2_{(1)}+\chi^2_{(2)}+\cdots+\chi^2_{(b-1)}$$

と表わすことができる．このことからも，U検定は p_{1j}/p_{0j} の j に関する線形な変動のみに荷重を置いたきわめて指向性の強い検定であること，逆に χ^2 検定は，1次から $b-1$ 次までの変動に一様に荷重を置いた，いわゆる omnibus 検定で，臨床試験の優劣比較にはそぐわないことがわかる．これに対し χ^{*2} 検定は1次，2次，3次，… の変動に 6:2:1… の荷重を置いた統計量であり，主として1次的な変動に興味があるがそれに2次あるいは3次的な変動が混り込んだような場合にも高い検出力を持ち，実際的な検定法であると思われる．これらの性質はカテゴリの出現度数に四，五倍程度のアンバランスがあっても近似的に成立する（広津，1985）．

χ^{*2} 検定を実際に用いる場合には帰無仮説の下での分布を χ^2 分布の定数倍 $d\chi^2_\nu$ で近似する．定数 d，ν は χ^{*2} と $d\chi^2_\nu$ で平均と分散が一致するように定める．定数 d，ν を求める公式は広津（1983, 84b）に与えてあるが若干面倒との批評も頂戴した．しかし，広津（1984b）を $b=3$，4の場合について書き下すと容易に一般ルールに気付かれると思う．たとえば $b=3$，4について d を書き下すと次のようになる．

$b=3$：

$$d=1+\frac{2}{3-1}\left\{\frac{y_{.3}}{y_{.1}+y_{.2}}\left(\frac{y_{.1}}{y_{.2}+y_{.3}}\right)\right\}$$

$b=4$：

$$d=1+\frac{2}{4-1}\left\{\frac{y_{.3}+y_{.4}}{y_{.1}+y_{.2}}\left(\frac{y_{.1}}{y_{.2}+y_{.3}+y_{.4}}\right)+\frac{y_{.4}}{y_{.1}+y_{.2}+y_{.3}}\left(\frac{y_{.1}}{y_{.2}+y_{.3}+y_{.4}}+\frac{y_{.1}+y_{.2}}{y_{.3}+y_{.4}}\right)\right\}$$

ただし，$y_{.j}$ は第 j カテゴリの総和である．まず各分数式の分子は $y_{..}(=y_{.1}+\cdots+y_{.b})$ から分母を引き去ったものとして自動的に得られるから分母だけに注目すればよい．次に{ }内で（ ）の外にある項の分母は $y_{.1}+y_{.2}$ からはじめて順番に $y_{.1}+\cdots+y_{.b-1}$ までが現われる．この各項（たとえば $y_{.1}+\cdots+y_{.j}$）に対し，（ ）内には分母が $y_{.2}+\cdots+y_{.b}$，…，$y_{.j}+\cdots+y_{.b}$ である $j-1$ 個の項が現れる．最後に ν は

$$\nu=(b-1)/d$$

として求めることができる．

以上，順序カテゴリカル・データに基づく2剤比較のための代表的な検定法についてその特徴付けを述べた．一組の試験結果に，これら特徴の異なる検定法を複数適用すると，いわゆる後知恵の多重推測となり有意水準が過大になってしまう．第Ⅲ相の臨床試験では厳密な検証が目的となるから，Ⅱ相までの試験結果をよく検討した上で，これら検定の特徴をよく考慮し事前にどれか一つを選択しておかなければならない．

3.2）多剤比較について

臨床試験は必ずしも前節で扱ったような治験薬対対照薬の2剤比較ばかりとは限らない．たとえば a 通りの薬剤に関して治験が行われたとすると薬剤間の比較の自由度が $a-1$ になり，どのような差に注目するかによって検定法を選択する余地が生ずる．たとえば連続量データの場合に a 個の正規母集団の平均を比較検定する方法としてよく見かけるだけでも，omnibus な一様性検定である F検定，対比較統計量の最大値を用いる Studentized Range 法（Tukey の多重比較法），あるいは標準との比較を行う Dunnet の多重比較法などがある（Scheffé の

多重比較法はF検定と同値である). これが順序カテゴリカル・データ ($a\times b$ 順序分割表) となると, このそれぞれに順序カテゴリをどう扱うかによって前節述べたような諸法の組み合わせが考えられる.

たとえば, omnibus なF検定に対応するものとして, 各薬剤の得た順位和を用いる Kruscal-Wallis 検定 ($a=2$ のときは両側U検定にほかならない), $b-1$ 通りの断点でプールして構成した $a\times 2$ 分割表の χ^2 の和 $\chi^{*2}=\chi_1^2+\cdots+\chi_{b-1}^2$ に基づく χ^{*2} 検定, 順序を無視した一様性の χ^2 検定, そして χ^{*2} の最大成分 $\max(\chi_1^2, \cdots, \chi_{b-1})$ に基づく $\max\chi^2$ 検定などが考えられる.

第II相の臨床試験で検証を目的とする場合には, これら数多くの特性の異なる検定を有意水準の調整なしに重複して用いることは許されないから, 検証の目的を明確にした上で再び各検定の特徴をよく考慮し, 事前に検定法を一つ選択しておかなければならない. 有意水準 α の二種類の検定を重複適用すれば, 全体としての有意水準は 2α とはいわない(二種の検定が一般に独立ではないから)までもそれに近くなってしまうことは明らかである.

順序カテゴリの取り扱い方の違いによる検定の特徴付けについてはすでに前節で詳しく述べたことだし, あらゆる組み合わせについて論じるのはいたずらに話を複雑化するだけだから, ここでは a 個の正規母集団の平均の比較について考える. この場合の検定の特徴付けは, ほとんどそのまま順序カテゴリカル・データの場合にも有効である.

実は正規母集団の群比較については以前 (1984 a) にも種々の検定の特徴付けを述べたのだが, それが一部で誤って一様性検定を支持するものと受けとられたようなので再度簡単に解説しておきたい.

いま a 通りの薬剤 $1, \cdots, a$ の特性値が正規分布 $N(\mu_1, \sigma^2), \cdots, N(\mu_a, \sigma^2)$ に従っているとする. 各薬剤に対しデータが $y_{i1}, \cdots, y_{in} (i=1, \cdots, a)$ のように得られているとする. 繰り返し数は各薬剤について必ずしも揃っている必要はないが, ここでは便宜上共通に n としておく.

さて, a 通りの薬剤に関する一様性の帰無仮説

$$H_0: \mu_1=\cdots=\mu_a \tag{3.4}$$

の検定法として標準的なテキストに書かれているのはF検定である. しかしながら筆者は, F検定のようなOmnibus 検定は臨床試験の解析にはそぐわないと考えている.

F検定は $\mu_1, \cdots, \mu_a$ の間のいかなる差に関してもまったく同等の検出力を持っている. したがって対の差 $\mu_i-\mu_{i'}(i, i'=1, \cdots, a;\ i\neq i')$ の他 $(\mu_1+\mu_2)/2-\mu_3$ や, $0.1\mu_1+0.6\mu_3-0.2\mu_2-0.5\mu_4$ のような奇妙なものまで, 係数の和が0になるような μ_i の線形結合 (これは対比と呼ばれる) なら何でも同じように興味のあるとき用いるべき検定である. しかるに臨床試験で興味があるのは一般に2薬剤間の差

$$\mu_i-\mu_{i'}(i, i'=1, \cdots, a;\ i\neq i') \tag{3.5}$$

とか, あるいは薬剤1が対照薬で他の $a-1$ 剤が治験薬の場合には治験薬と対照薬の差

$$\mu_i-\mu_1(i=2, \cdots, a) \tag{3.6}$$

などであろう. このように興味の対象が絞られれば, 意味のない対比に対する検出力を犠牲にすることによって, 真に興味のある対比に対する検出力を上げることができる. このような考え方で構成された検定を omnibus 検定に対して指向性検定と呼ぶ.

対比較 (3.5) を対象とした検定 (Studentized Range または Tukey の多重比較法) の棄却域は次で与えられる.

$$R_T: \max_{i, i'}\left\{\frac{\sqrt{n}\,|\bar{y}_{i\cdot}-\bar{y}_{i'\cdot}|}{\hat{\sigma}}\right\}>q_\alpha\{a, a(n-1)\} \tag{3.7}$$

ただし, max はすべての i, i' の組み合わせでの最大値を意味し, $\bar{y}_{i\cdot}$ は $y_{i1}, \cdots, y_{in}$ の平均, $\hat{\sigma}$ は全データから計算した不偏分散

$$\hat{\sigma}^2=\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})/\{a(n-1)\}$$

の平均根である.

対照薬との差 (3.6) を対象とする Dunnet の検定の棄却域は次で与えられる.

$$R_D: \max_{i}\left\{\sqrt{\frac{n}{2}}\cdot\frac{|\bar{y}_{i\cdot}-\bar{y}_{1\cdot}|}{\hat{\sigma}}\right\}>d_\alpha''\{a-1,\ a(n-1)\} \tag{3.8}$$

これらの検定のための棄却限界値 q_α, d_α'' などは数表で与えられる (たとえば広津, 1983, 付表9および11).

ここで $a=3$, $n=3$, $\alpha=0.05$ の場合を例にとって各検定の棄却域を図示して見よう. 普通の直交座標に描くために y を次式で定義される z に変換する.

$$\begin{cases} z_1=\sqrt{\dfrac{3}{6}}\cdot\dfrac{2\bar{y}_{3\cdot}-\bar{y}_{1\cdot}-\bar{y}_{2\cdot}}{\hat{\sigma}} \\ z_2=\sqrt{\dfrac{3}{2}}\cdot\dfrac{\bar{y}_{2\cdot}-\bar{y}_{1\cdot}}{\hat{\sigma}} \end{cases}$$

妙な係数は規準化のための定数である.

この変換によりF検定の棄却域は次のようになる.

$$R_F: n\sum_{i=1}^{3}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})/\{(3-1)\hat{\sigma}^2 = \frac{z_1^2+z_2^2}{2}>F_{0.05}(2, 6)=5.143 \tag{3.10}$$

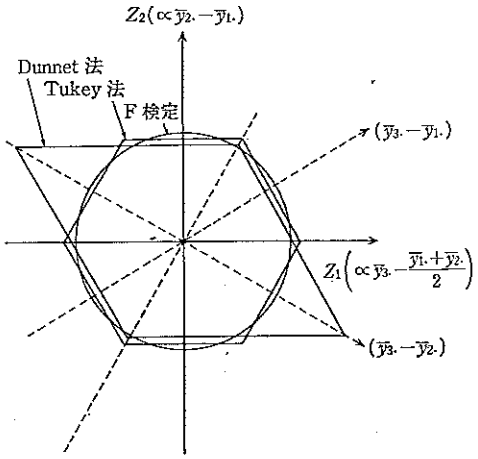

Fig. 5 三種の検定の図示
(a=3, n=3, α=0.05)

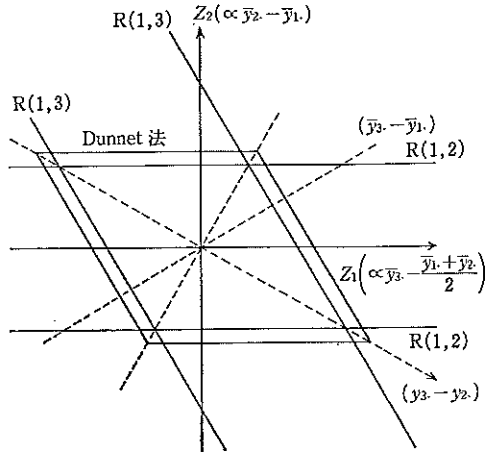

Fig. 6 **Dunnet 法と検定の比較**

同様に Tukey 法, Dunnet 法の棄却域はそれぞれ次のようになる.

$$R_T:\ \max\left\{\sqrt{2}\,|z_2|,\ \frac{|\sqrt{3}z_1+z_2|}{\sqrt{2}},\ \frac{|\sqrt{3}z_1-z_2|}{\sqrt{2}}\right\} > q_{0.05}(3,\ 6)=4.34 \qquad (3.11)$$

$$R_D:\ \max\left\{|z_2|,\ \frac{|\sqrt{3}z_1+z_2|}{2}\right\} > d_{0.05}{}''(2,\ 6) = 2.86 \qquad (3.12)$$

この結果を図示したのが Fig. 5 である.

z_1, z_2 の, 帰無仮説 H_0(3.4)の下での同時分布は原点に中心のある富士山のような形をしており, いずれの棄却域についてもその上の体積 (有意水準) は 0.05 である.

F検定の棄却域はこの図で円の外部であり, 帰無仮説がどの方向にくずれても同じように検出することが明らかである. Tukey 法では予期したように, 対比較 ($\mu_i-\mu_{i'}$, $i,\ i'=1,\ 2,\ 3;\ i\neq i'$) に対してF検定より検出力が高くなる. Dunnet 法は治験薬同士の差 $\mu_3-\mu_2$ の比較を犠牲にすることにより興味のある $\mu_2-\mu_1$, $\mu_3-\mu_1$ に対する検出力を上げている. このように, 治験薬と対照薬の比較に主たる興味があるときに, F検定や Tukey 法を用いると相当な検出力の低下を免れない. このことはとりあげた薬剤数 a が大きいときに, より顕著である.

次に Dunnet 法を薬剤2と1を比較するt検定の棄却域 $R(1,2)$ および薬剤3と1を比較するt検定の棄却域 $R(1,3)$ と較べてみよう.

帰無仮説 $H(1,\ 2):\mu_1=\mu_2$ に対する両側t検定の棄却域は

$$R(1,2):\sqrt{\frac{n}{2}}\left|\frac{\bar{y}_{1\cdot}-\bar{y}_{2\cdot}}{\hat{\sigma}}\right| = |z_2| > t_{0.05/2}(6)=2.447 \qquad (3.13)$$

である. 同様に $H(1,\ 3):\mu_1=\mu_3$ に対する棄却域は

$$R(1,3):\sqrt{\frac{n}{2}}\frac{|\bar{y}_{1\cdot}-\bar{y}_{3\cdot}|}{\hat{\sigma}} = \frac{|\sqrt{3}z_1+z_2|}{2} > 2.447 \qquad (3.14)$$

で与えられる. ただし, Dunnet 法と同一の図に表わすために不偏分散としては全データから計算される $\hat{\sigma}^2$ を使うものとした. これらの棄却域を図示したのが Fig. 6 である. Fig. 6 はもし薬剤2と1の比較だけに興味があるならば, 当然ながら Dunnet 法はt検定 $R(1,2)$ より検出力が劣るし, また薬剤3と1の比較だけなら検定 $R(1,3)$ より劣ることを示している. つまり Dunnet 法といえども事前に (II相までの試験で) 治験薬を一つに込め絞りなかったつけを払わされている.

さらにこのことは, 検定で有意差が出なかったことをもって同等性を主張する風潮の下では単なる検出力の低下では済まされない. たとえば対照薬との対比較でかろうじて有意差が検出される程度に劣った治験薬を二つ同時に取り入れた臨床試験で Dunnet 法を用いると, 検出力の低下のために有意差が検出できず誤った同等性が主張される危険性がある. ましてや omnibus なF検定を行うと, この危険性は一層増大する. この危険性は薬剤の数を増やすほど大きくなり, 平凡な薬剤を数多く含むと, その中に相当に劣った治験薬が紛れこんでいてもまず有意差が検出されることはなくなるだろう. それゆえできることなら第II相までの試験をよく行い, 検証を目的とする第III相は2剤試験であることが望ましい. とく

Table 16　3剤比較のデータ例

(大泉・他(1986)から引用，ただし，原著 Table 9 から不明の2例を省いた)

カテゴリ	無効	やや有効	有効	著効	計
1. AMPC	3	8	30	22	63
2. S 6472	8	9	29	11	57
3. CCL	2	11	33	17	63

に1群当りのサンプル・サイズではなく，総数に制限があるようなら，3剤比較と2剤比較では1群当りのサンプルサイズが1.5倍も違うことになる．Fig. 6 の上でいえば $R(1,2)$ や $R(1,3)$ はさらに $\sqrt{2/3}=0.816$ 倍に縮小され，検出力の違いは相当になる．

やや話が抽象的に流れたので一つ実例をあげて注意を喚起したいと思う．

Table 16 で薬剤1が対照薬，薬剤2および3が治験薬である．これは順序カテゴリカル・データであるが話の筋は正規分布モデルの場合と同様である．ここでは簡単のため順序カテゴリの処理は順位和法を用いることとし，3薬剤間の一様性検定には Kruscal-Wallis 検定を用いてみる．

Table 16 のデータについて，まずそれぞれの治験薬と対照薬の間で検定を行って見よう．まず，S 6472 と AMPC の間では規準化した Wilcoxon-Mann-Whitney 統計量は

$$|U(1,2)|=2.182$$

となり，両側5％有意である．すなわち，治験薬 S 6472 は対照薬 AMPC より劣っている．次に，CCL と AMPC の間では規準化した Wilcoxon 統計量として

$$|U(1,3)|=0.864$$

が得られる．これは両側5％検定で有意ではない．

次に3薬剤間についての一様性検定を試みよう．この場合 Kruscal-Wallis の検定統計量は

$$KW=5.372 \text{ (自由度}=2) \qquad (3.15)$$

となる．自由度2のカイ二乗分布の5％点は5.99だからこれは有意ではない．実際 (3.15) 式の有意確率は7％弱である．この結果対照薬との2剤比較なら棄却できたはずの治験薬が，他の治験薬とこみで3剤比較が行われたために棄却できないことが生じ得る．これは少し話がおかしくないだろうか．

このため，3剤の同時試験をやっていても，従来よく行われていたように，各治験薬対対照薬の2剤比較を行えばよいではないかとの議論が出てくるかも知れない．しかし，それはやはり誤りであって有意水準が過大になり言い過ぎが生じるのは明らかである．Fig. 6 の場合に戻っていうなら，有意水準0.05の対比較を2回行ったときは，結局小さい菱形の外部を棄却域としていることになる．大きい菱形の外部を棄却域とする検定の有意水準がちょうど0.05なのだから，小さい菱形を用いれば有意水準は7, 8％にはなるだろう．有意水準が狂ってしまえば元も子もないのであって，必要なことは，あくまで検定の有意水準は確保した上でなるべく検出力を高めることである．そのための一番簡単な方法はII相試験までで治験薬を絞りこみ，標準的薬剤を対照として2剤比較を行うことなのだが，そうはいっても実際の情況はもっと複雑でどうしても多剤比較の要請は出てくることと思う．

すでに述べたように多剤試験の場合は比較の自由度が複数になるので，あらかじめ検証したい仮説をきちんと規定しておくことが大切である．それは一つには多剤試験の構成にもよることである．たとえば同じ3剤試験でも2治験薬，1対照薬の場合，1治験薬2対照薬の場合，そして治験薬，対照薬（アクティブ），プラセボの3剤比較の場合ではおのずと臨床試験の目的，従って統計的帰無仮説，対立仮説は違ってくるだろう．これらあらゆる場合を想定して述べることは繁雑なので以下では最もポピュラーであると思われる2治験薬，1対照薬の場合についてのみ考える．この2治験薬とは，同一薬剤で dose level を変える場合とか，投与経路を変える場合等を含んでいる．

この場合目的となるのは当然ながら，対照薬と較べて同等以上の治験薬はなるべく採択し，劣っている治験薬はなるべく棄却したいということだろう．この場合，対照薬の特性値の母平均を μ_1，治験薬 2, 3 の母平均をそれぞれ μ_2, μ_3 とするとき，仮説

$$H_{12}: \mu_1=\mu_2 \qquad (3.16)$$

$$H_{13}: \mu_1=\mu_3 \qquad (3.17)$$

に対する Dunnet の多重比較法を適用するのが最も普通の考え方だが，それに変えて次のような信頼方式を採ることが考えられる．すなわち，次のような仮説，

$$H_0: \mu_1=\mu_2=\mu_3 \qquad (3.18)$$

$$H_{12}': \mu_2=\mu_1 \neq \mu_3 \qquad (3.19)$$

$$H_{13}': \mu_3=\mu_1 \neq \mu_2 \qquad (3.20)$$

$$H: \mu_2 \neq \mu_1 \neq \mu_3 \qquad (3.21)$$

のそれぞれに対して有意水準 α の検定を行い，棄却されなかった仮説の集合を採択するという陳述の信頼率は $1-\alpha$ 以上であることを用いる．

ここで H_0 の検定には仮説 (3.16), (3.17) を対象とした Dunnet の両側検定，H_{12}' に対しては (3.16) に

対する両側t検定，H_{13}' に対しては(3.17)に対する両側t検定を考える．ただし，いずれも標準偏差の推定量としては全群からの（自由度 $3(n-1)$）ものを用いる．仮説H(3.21)に対する棄却域としては Fig. 6 の上で，原点を中心としたきわめて小さい菱形が考えられるが，それを無視してもあまり大きな影響はないのでここの議論からは省く．そうすると，信頼率 $1-\alpha$ で次のような陳述が可能である．

簡単のため

$$t_{12}=\sqrt{\frac{n}{2}}\frac{\bar{y}_{2\cdot}-\bar{y}_{1\cdot}}{\hat{\sigma}},\quad t_{13}=\sqrt{\frac{n}{2}}\frac{\bar{y}_{3\cdot}-\bar{y}_{1\cdot}}{\hat{\sigma}}$$

とおき，$|t_{13}|>|t_{12}|$ とする．

まず，$|t_{13}|<d_\alpha''\{3(n-1)\}$ なら H_0 が採択されるから，他の仮説の採否とは無関係に治験薬2，3と対照薬の間に有意差はないということにする．それはたとえ H_0 に加えて H_{12}' が採択されてもそれは μ_3 と μ_1 は等しいか，等しくないかどちらの可能性も否定できないことを意味するだけだからである．次に $|t_{13}|>d_\alpha''\{3(n-1)\}$ なら H_0 は棄却され，また当然 $|t_{13}|>t_\alpha\{3(n-1)\}$ となり H_{13}' も棄却される．そこで H_{12}' だけについて考えればよいがそれは $|t_{12}|\leq t_{\alpha/2}\{3n(-1)\}$ なら採択，$|t_{12}|>t_{\alpha/2}\{(n-1)\}$ なら，棄却される．つまり $|t_{13}|>d_\alpha''\{3(n-1)\}$ のとき $|t_{12}|>t_{\alpha/2}\{3(n-1)\}$ なら，両治験薬ともに対照薬との間に有意差が示され，$|t_{12}|\leq t_{\alpha/2}\{3(n-1)\}$ なら，治験薬3と対照薬の間には有意差があるが，治験薬2と対照薬の間には有意差がないと結論すればよい．以上要するに $|t_{12}|$ と $|t_{13}|$ の大きい方が Dunnet の意味で有意でないとき治験薬と対照薬の間に有意差は無いとし，両者のうち大きい方が Dunnet の有意点を越えるとき，小さい方が両側t検定の棄却限界値を越えれば両治験薬は対照薬との間に有意差があるとし，もし越えなければ大きい方だけが対照薬との間に有意差があるということにすればよい．

ここで Table 16 の例にこの方法を適用して見よう．まず，大きい方の U 値，

$$|U(1,2)|=2.182 \qquad (3.22)$$

は Dunnet の規準で評価しなければならない．ただし，5%点が数表で与えられているのは正規分布比較で繰返し数が等しい場合だから，ここでは(3.22)式の有意確率を(3.23)式に従って直接評価する．

$$\begin{aligned}&Pr\{\max\{|U(1,2)|,|U(1,3)|\}<2.182\}\\&=Pr\{|U(1,2)|>2.182\}+Pr\{|U(1,3)|>2.182\}\\&-Pr\{U(1,2)>2.182,\ U(1,3)>2.182\}\times 2\\&-Pr\{U(1,2)>2.182,\ U(1,3)<-2.182\}\times 2\end{aligned}\qquad (3.23)$$

(3.23)式の1項，第2項はそれぞれ標準正規分布の上側確率の計算によって

$$2\times 0.01456\fallingdotseq 0.0291$$

と得られる．次に $U(1,2)$ と $U(1,3)$ の相関係数が

$$\sqrt{\frac{57\times 63}{(57+63)(63+63)}}=0.48734$$

となる（広津，1983参照）ことから，第3項，第4項はそれぞれ 0.00208×2，0.00008×2 となる．以上から(3.23)式の値は

$$0.0291\times 2-0.00208\times 2-0.00008\times 2\fallingdotseq 0.0539$$

となる．すなわち $|U(1,2)|=2.182$ の有意確率は約5.4%である．これは先ほどの Kruscal-Wallis 統計量の有意確率が7%弱であったのに較べれば大きな改善ではあるが，あくまで $\alpha=0.05$ に固執するならば有意差は示されないことになる．従って $|U(1,3)|=0.864$ を標準正規分布を参照して検定するステップには進めないことになる．

ただし，ここで有意差が無いと称するのは，単に異なっているとする十分な証拠がないというだけであって，積極的な同等性を意味するものでないことはもちろんである．したがって，有意差が示されないことをもって同等とみなすというのは誤りであるが，現在，その誤った風潮にあるとすれば，劣った治験薬を排除するために検定の検出力を確保しておかなければならない．

椿（1985）は3剤比較の Dunnet 法で，2剤比較のt検定と同等の検出力を保持するためのサンプル・サイズを計算し，1群当りのサンプル・サイズを約14%減らしてよいことを述べている．一方，藤田他（1986）は劣った薬剤を排除するためには個々の対比較で2剤比較と同等の検出力を保持する必要があるとの観点から，椿（1985）とは逆に3剤比較の場合に1群当り約1.27倍のサンプル・サイズが必要であると論じている．確かに有意差が示されないことをもって同等とみなす風潮の下では $|t_{12}|$ と $|t_{13}|$ の大きい方だけの検出を対象とした椿（1985）の考え方は optimistic に過ぎるが，逆に $|t_{12}|$ と $|t_{13}|$ の小さい方を Dunnet の規準と比較するとの考えに基づいた藤田他（1986）の結論は pessimistic に過ぎると思われる．ここで述べた信頼方式によれば，$|t_{12}|$ と $|t_{13}|$ の小さい方の検定は2剤比較のt検定に帰着するから，3剤比較の場合も1群当りのサンプル・サイズは2剤比較と同様にすればよいことになる．2剤比較の例数設計は直観的にも分りやすく，よく行われてもいるので，この結果は有用と思う．

さて，同等性検証として d 以上劣っていないことを証明するという2節の考え方はこの場合には次のように拡

張される．すなわち，仮説(3.18)～(3.20) の代りに，

$H_0(\Delta)$：$\mu_2+\Delta<\mu_1$，$\mu_3+\Delta<\mu_1$

$H_{12}(\Delta)$：$\mu_2+\Delta<\mu_1$

$H_{13}(\Delta)$：$\mu_3+\Delta<\mu_1$

を考える．統計量としては

$$t_{12}(\Delta)=\sqrt{\frac{n}{2}}\frac{\bar{y}_{2\cdot}+\Delta-\bar{y}_{1\cdot}}{\hat{\sigma}},$$

$$t_{13}(\Delta)=\sqrt{\frac{n}{2}}\frac{\bar{y}_{3\cdot}+\Delta-\bar{y}_{1\cdot}}{\hat{\sigma}}$$

を採り，$H_0(\Delta)$ の検定には Dunnet の片側検定，$H_{12}(\Delta)$, $H_{13}(\Delta)$ の検定には片側 t 検定を用いる．そして $\max\{t_{12}(\Delta), t_{13}(\Delta)\}$ が Dunnet の意味で有意であり，かつ $\min\{t_{12}(\Delta), t_{13}(\Delta)\}$ が t 検定で有意なら，両治験薬ともに対照薬より Δ 以上は劣らないことがいえたとする．$\max\{t_{12}(\Delta), t_{13}(\Delta)\}$ が Dunnet の意味で有意で，かつ $\min\{t_{12}(\Delta), t_{13}(\Delta)\}$ が t 検定で有意でないなら $t_{12}(\Delta)$ と $t_{13}(\Delta)$ のうち大きい方の治験薬についてのみ，対照薬より Δ 以上劣っていないことが証明されたとすればよい．

3.3) まとめ

データが著効，有効，無効といった順序カテゴリの発現頻度の場合には順序カテゴリをどう扱うかによって Wilcoxon-Mann-Whitney 検定，累積カイ二乗検定，max χ^2 検定などが考えられる．これらの特徴は 3.1) 節で詳しく述べた．第Ⅱ相試験までで薬の特徴がよくつかめているときには，その特徴に応じて Wilcoxon-Mann-Whitney 検定や max χ^2 検定を選ぶことができる．その反対に特徴が十分つかみきれていないときには，優劣差を表す幅広い対立仮説の下で高い検出力を保持する累積カイ二乗検定が推薦できる．第Ⅱ相試験までの解析なら，これら特性の異なる検定を種々適用して探索を行ってもよいが，検証を目的とする第Ⅲ相では，有意水準を保つためにあらかじめ検定法を規定しておくことが大切である．

第Ⅲ相試験はできることなら治験薬対標準薬の2剤比較が望ましい．ことさら多剤比較を行って1群当りのサンプル・サイズを減らすより，2剤比較にして十分なサンプル・サイズを確保することが望まれる．どうしても多剤試験を行うときには，その目的を明確にし，それに応じた検定手法を選択しておかなければならない．多剤試験の場合は（ちょうど順序カテゴリの扱い方が種々あるように）比較の自由度が複数なので，事後的に興味の対象を絞る後知恵解析に陥らないためである．

2治験薬と1対照薬を比較する3剤試験では Dunnet の多重比較法を用いることを原則とする．ただし，大きい方の t 値が Dunnet の意味で有意であるときに小さい方の値の有意性をいうのには普通の t 検定を行ってよい．従って検出力確保のための1群当りのサンプル・サイズは2剤比較で要求される1群当りのサンプル・サイズと等しくとればよい．この信頼方式の考え方は，Δ 以上劣っていないことを証明するというネガティブ・トライアルの考え方にも適用される．

臨床試験における多重性は 3.1) 節，3.2) 節で述べた場合の他，事後的な層別解析，経時的に繰り返しとられた，いわゆるリピーテッド・メジャメントの解析，複数項目の総合解析などにおいても問題となる．これらについては機会をあらためて議論したい．

文　献

1) 江島　昭・他：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, 1982.
2) 大泉耕太郎・他：細菌性肺炎に対する S 6472, Cefaclor と Amoxicillin の二重盲検法による臨床評価の比較．*Japanese J. Antibiotics*, 39(2)：853-886, 1986.
3) 椿　広計：多重比較の検出力について，標準のある3群比較の場合．品質管理学会第15回年次大回要旨：29-32, 1985.
4) 広津千尋：離散データ解析．教育出版，東京，1982.
5) 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.
6) 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.
7) 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, 1984 b.
8) 広津千尋：順序分割表における残差分析．応用統計学，14(2)：61-67, 1985.
9) 広津千尋：臨床試験における統計的諸問題 (1)．臨床評価，14：467-475, 1986.
10) 藤田利治，椿　広計：臨床試験の結果の検定法による相違．臨床評価，13：601-611, 1985.
11) 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.
12) 森口繁一（編）：新編統計的方法．日本規格協会，東京，1976.
13) Dunnet, C.W. & Gent, M.: Significance testing to establish equivalence between treatments, with special referrence to data in the form of 2×2 tables. *Biometrics*, 33: 593-602, 1977.
14) Moses, L.E., Emerson, J.D. & Hosseini, H.: Analyzing data from ordered categories. *New Engl. J. Med.*, 311: 442-448, 1984.

15) Hirotsu, C.: Cumulative chi-squared statistic as a tool for testing goodness of fit. *Biometrika*, 73: 165-173, 1986.
16) Sugiura, N. & Otake, M.: Approximate distribution of the maximum of $(c-1)$ χ^2 statistics (2×2) derived from $2\times c$ contingency table. *Communication in statistics*, 1: 9-16, 1973.

* * *