The Society for Bioscience and Bioengineering, Japan

〔生物工学会誌　第85巻　第5号　227–232. 2007〕

講座 ケモメトリックス入門

第2回　多重比較法入門

高木　達也

多重比較はなぜ必要か

前世紀末（と言ってもわずか10年ほど前のことなのだが），「ノストラダムスの大予言」なるものが流行した．中身を読んでも，ふわっとしたことしか書いていないので何とでも解釈できるのだが，解釈の仕方によっては，ことごとくその後の歴史を言い当てているように思えるため，「1999年，7か月，空から恐怖の大王が来るだろう」という詩の一部から，次の千年紀を迎える前に，地球には大厄災が訪れるのではないかと話題になった．現金なもので，2000年に入るとノストラダムス関連の書籍はぱったりと売れなくなったそうであるが，それでは，それ以前の予言はなぜ当たっているように思えたのだろうか．ノストラダムスは1999年だけはずしてしまったのだろうか．

ミシェル・ノストラダムスは，16世紀前半から中葉のマニエリスムの時代を生きた，元来は医師であった．パルミジャニーノ，アノーロ・ブロンツィーノ，ジャコポ・ダ・ポントルモ，ロッソ・フィオレンティーノらが画家として活躍し，ウィリアム・シェークスピア，ジョン・ダンが小説家として，ジョヴァンニ・ピエルルイジ・ダ・パレストリーナやトーマス・タリスが作曲家として活躍した時代，象徴，隠喩，暗示，寓意，幻想に充ち満ちた文化は，その背景に「ストックホルムの血浴」や，「ドイツ農民戦争」，「ローマの劫掠」など，ルネサンスの響きとは反対に，宗教改革，宇宙観の転換など，価値観が音を立てて崩壊してゆく時代があったと思われる．このような未来が見えない時代には予言者が多数現れる．ノストラダムスより少し前，フィレンツェを震撼させた怪僧ジロラモ・サヴォナローラも多数の予言を行ったが，醜態を演じたとして火刑に処せられている．

さて話を元に戻して，ノストラダムスの予言が当たった理由を考えてみよう．ノストラダムスのように曖昧模糊とした予言だと，拡大解釈をすれば，かなりの確率で「当たった」と言える．しかも予言書を書いたのはノストラダムスだけではあるまい．現にサヴォナローラや聖マラキの予言？など，有名なものだけでも両の指に余る．失われたものも含めれば，数え切れないほどの予言書があったと思われる．これらの中で，どう拡大解釈しても当たらなかったものは，サヴォナローラではないが廃棄されたのだろう．そして，聖マラキの予言など「当たったと言えなくもない」ものが現在も「生きている」と考えると，ノストラダムスは単に「ついていた」だけと言うことになる．ノストラダムスにとって不幸なことに，「恐怖の大王」としか書いていないものを，世界の終焉と結びつけられてしまったがために廃棄される側に回ってしまった．要は，1つ1つの予言書が残る確率は低くても，たくさんの予言書があれば一つくらいは残ってしまうことがあると言うところがミソである．宝くじだって億円単位で当たった方は読者の中にもほとんどいないと思われるが，日本のどこかには存在するはずである．

「当たり前のことを言うな」とお叱りかも知れないが，これが現実のデータの処理となると，とたんに忘れられてしまうから筆者は不思議で仕方がない．しつこいかもしれないが，ノストラダムスの予言と違って，もう少し定量的に語れる例を出してみたい．

皆さんの小・中・高校時代を思い出して頂きたい．皆さんのクラスに，誕生日がまったく同じクラスメイトはいただろうか？「いた」という方は，おそらく少数だろう．誕生日がまったく一致するという現象は，それほど頻度の高いものではない．では，皆さんご自身とは限らず，誕生日がまったく同じ組み合わせのクラスメイトはいただろうか？「それだって頻度の高い現象ではないから，まずはあり得ないだろう」と思われた方は，すでに本回の主題でもある，「多重比較」の罠にはまっている可能性がる．よく考えてみよう．確かに，「自分と同じ誕生日のクラスメイトがいない」確率は，$p_1 = 1 - \frac{1}{365} = \frac{364}{365}$ で，

著者紹介　大阪大学大学院薬学研究科（教授）　E-mail: ttakagi@phs.osaka-u.ac.jp

NII-Electronic Library Service

The Society for Bioscience and Bioengineering, Japan

まあ，たいていの場合，そんなクラスメイトはいないと考えていい．しかし，これがクラス40人すべての組み合わせとなると，1人（もしくは2人）ではあり得なかったことも，${}_{40}C_2 = 780$ 通りの組み合わせを考える必要があり，そう話は単純ではなくなってくる．計算してみると，40人中，同じ誕生日の組み合わせが存在する確率は，「存在しない」確率を1から差し引けばいいので，

$$1-\prod_{i=1}^{40}\frac{365-i}{365}=1-0.096848=0.903152$$

と，90%以上になる．一人ではありえない組み合わせも，780もの組み合わせがあれば，たいてい，どこかで見つかることになる[1]．要は「下手な鉄砲数打ちゃ当たる」ということである．アドルフ・ヒトラーは「大衆は馬鹿だ…嘘も百回繰り返せば，信じ込む」と言ったそうであるが，確率論的に考えるならば，残念ながらこれは正しい．したがって，私たちはヒトラーのような奴に騙されないようにしなければならない．

「そんなことは，教えてもらわなくても解っている」とお思いだろうが，実は，解っているつもりでも，いざ統計的有意性検定に利用しようとなると，忘れられていることが少なくないので困っている．多重比較なんて，しょせん「下手な鉄砲数打ちゃ当たる」ことを想定して検定を行おうというものなので，たいしたことはないのだが，気がつくか気がつかないか，気づいたときに正しい補正手段が選べるかどうかが重要になってくる．しかも高名な統計ソフトに，使ってはいけない，または使うときに注意する必要のある手法が昔のまま残っていたりするので困ったものである．以下，いま少し正確に，多重比較法を見ていこう．

多重比較が必要な有意性検定

今，ある5種類の製剤の薬理活性を互いに比較すると考える．実際に測定してみたところ，表1のような結果が得られた．これら5種類の製剤の薬理活性は，互いに統計的に有意に異なっていると言えるだろうか？…というような問題に出くわす，あるいは出くわされた場合は少なくないと思う．この場合，どのような計算をしましたかと聞くと，ちょっと前までは「Studentのt検定をA－B，A－C，…，D－Eと繰り返して行って，5%水準で有意なら＊マーク1つ，1%水準で有意なら，＊マークを2つつけてグラフに表示しました．」（図1）という答えが，当然のごとく返ってきた．さすがに最近そんなことはなくなってきたが，今でもなぜこれではいけないのか，と聞かれることはしばしばある．筆者はその都度，前記のような話をさせて頂くのだが，40人780通りの組み合わせだと簡単に直感的に理解できることも，5種類になると人間，簡単に図1のような誤りを犯していたのだから，心理学的な興味をそそられる問題でもある．

表1．5種類の製剤の薬理活性値

	A	B	C	D	E
1	0.644	0.732	0.471	0.879	1.857
2	0.342	0.854	0.540	1.590	1.502
3	0.214	0.297	0.622	1.203	1.572
4	0.245	0.472	1.256	1.477	1.748
5	0.892	0.648	0.424	0.969	1.661
6	0.654	0.642	0.862	1.056	1.708
7	0.740	1.073	0.739	0.606	1.109
8	0.015	0.335	0.598	1.397	1.741
9	0.136	0.530	1.111	1.338	1.217
10	0.025	0.967	1.056	0.887	1.561

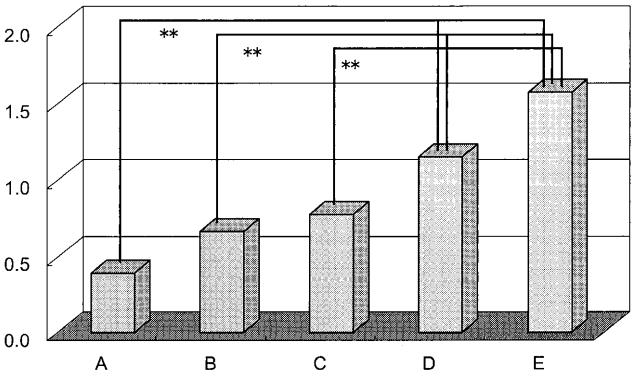


図1．**: $p<0.01$ by Student's t-test …ではいけない．

5種類の医薬品から2つを選ぶ組み合わせは，${}_5C_2 = 10$ 通りある．780通りに比べると数は少ないが，1でないという意味では定性的には大差ない．偶然に有意になる確率は，Studentのt検定で計算されるより遙かに大きくなる．多重比較とは，これを1%なら1%になるように，p 値を補正するなど，いろんなことをして，有意水準を当初の予定通りに保とうとする方法を指す．多重比較法の種類別に，簡単に見ていきたい．その前に今一度強調しておく．「多くのものを同時に比較し，有意差があることを主張したい場合，Studentのt検定を繰り返し用いてはいけない．」

p 値を直接補正する方法

p 値を直接補正する方法で最も簡単でしばしば用いられるのはBonferroni法である．名前は何か難しそうだが何のことはない，公称の有意水準 α（死守したい有意水準）を，検定の回数 n（表1のケースでは，10）で割るだ

NII-Electronic Library Service

The Society for Bioscience and Bioengineering, Japan

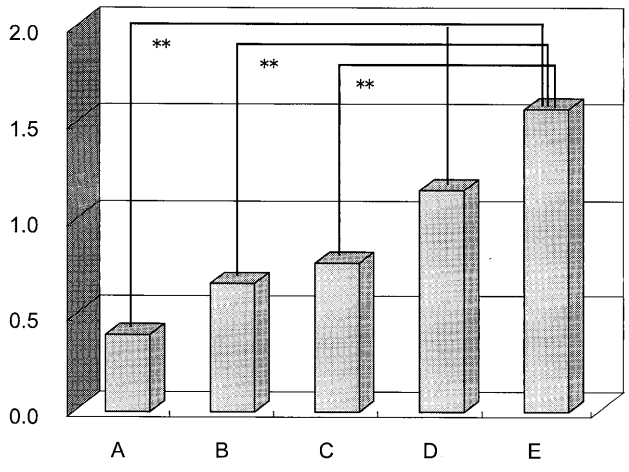


図2. Bonferroni法

けである.

$$\alpha' = \alpha/n \qquad (1)$$

これを用いて，先述のデータを再度有意性検定してみればどうなるか，図2がその回答である.

図1では5通りあった有意な組み合わせが，4通りになっている．変わっているのは，B－D間の有意差がなくなったことであり，もともとStudentのt検定での危険率は0.00134であった．Bonferroni法では実際に検定に用いる有意水準が10分の1になるため，0.001以下でなければならず，残念ながらお引き取り願ったということになる.

Bonferroni法はきわめて簡単で，ほとんどすべてのケースで使用できるので，「何だ．多重比較ってこれだけの話か」と思われるかもしれないが，簡単な話には棘があるものである（ちょっと諺が間違っているような気がするが深く考えないで頂きたい）．このケースでは，高々10通りに過ぎないので，大きな問題とならないのだが，これが先の誕生日の話のように数百通りもあればどうなるだろう．仮に1000分の1になれば，0.000001で，このバリアーを越えて有意になるような差なら，そもそも検定などしなくとも一見して明らかに異なることが解りそうなものである．実は，Bonferroni法における実質の有意水準は「これだけ小さな値をクリアーして有意になるようなら間違いなく大丈夫ですよ．実際は，もう少し甘くてもいいんですけどね」という値で，平たく言えば「真の値」より小さすぎるのである．そこで，もっと「真の値」に近づけようと，統計学者がさまざまに苦労することになる.

なお，p値を直接調整する方法は，この他Sidack修正法などがあり，また当研究室で開発されたBaggingを用いる方法[2]もそのひとつであるが，Baggingの解説がこの後の予定なので，稿を改めたい.

パラメトリックな方法

真の値に近づくには，いくつかの前提を置くことができれば都合がいい．通常は，データの母集団分布が正規分布に従ってくれると大変ありがたい．これは，特別に正規分布が人を惑わす魔力を持っているためではなく，正規分布関数（ガウス関数）が種々便利な性質を持っていることも理由の一つである．この前提の下に，先のような多重比較を行うのに用いられるのがTukey法である．Tukey法など，母集団分布が一定の確率分布に従っているという前提で行う方法を，多重比較によらず，パラメトリック法と呼ぶ．Tukey法は典型的なパラメトリック多重比較法である．Tukey法では，この他すべての群で母分散は等しいと考える（Studentのt検定に似ている）．Tukey法では帰無仮説と対立仮説は以下のようになる.

$H0_{ij}: \mu_i = \mu_j$ (i, j=1,2, … , m ; i<j)

$H1_{ij}: \mu_i \neq \mu_j$

具体的な計算方法は以下のようになる.

1）（当たり前だが）前もって有意水準，αを決めておく.

2）それぞれの群（表1ならA-E）ごとに，平均m_iと分散V_iを求める.

3）誤差自由度f_eと誤差分散V_eを求める.

$$f_e = \left(\sum_{i=1}^{m} n_i\right) - m$$
$$V_e = \frac{1}{f_e}\left\{\sum_{i=1}^{m} V_i (n_i - 1)\right\} \qquad (2)$$

4）検定統計量，t_{ij}値を求める.

$$t_{ij} = \frac{|m_i - m_j|}{\left\{V_e\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right\}^{\frac{1}{2}}} \qquad (3)$$

5）Student化された範囲の上側100α%点（多重比較を取り扱った書物なら巻末に表があるはずである）を

$Q(m, fe; \alpha)$とすると，$t_{ij} \geq \frac{Q(m, f_e; \alpha)}{\sqrt{2}}$ なら帰無仮説，$H0_{ij}$は棄却され，それ以外なら棄却できない（通常の検定と異なるので，間違っても採択してはいけない．後述）.

この他，たとえば表1のAが従来薬でB－Eが新薬の場合などで，AとB－Eのみを比較したい場合（コントロールが存在する場合）などはDunnett法が用いられる．い

NII-Electronic Library Service

The Society for Bioscience and Bioengineering, Japan

ずれもパラメトリックな方法なので，母集団分布などに一定の前提が必要となるが，検出力は高い．

なお，しばしば間違われるのだが，たとえばAがコントロールで，B－Eが処置群だとする．B－Eの間に順序がある場合（たとえば処置後の時間経過や投与した薬品類の量など）はDunnett法ではなく，Williamsの方法が用いられるべきである．もっとも，筆者はWilliamsの方法を用いたことは一度もない．相関分析や回帰分析などを行うか，有意なトレンドの検定を行う方がbetterに思えるからであるが，いかがだろうか．

ノンパラメトリックな方法

パラメトリックな方法があるからには，ノンパラメトリックな方法も当然ながら存在する．手法の詳細は省略するが，パラメトリックな手法と対応させて表にしてみたのでご参照頂きたい．

ノンパラメトリックな方法は，母集団分布が正規分布に従っている必要は，もちろん，ない．なら全部ノンパラメトリックな方法を用いればいいではないかと思われるかもしれないが，ノンパラメトリックな方法は，計量データを順序データに変換して用いるため，どうしても情報の損失を防ぐことができない．このため，検出力はパラメトリックな方法に劣ることも多く，できるならばパラメトリックな方法を用いたいところである．しかしながら，母集団分布が正規分布に従っていることを保証できることはそう多くはないので，ジレンマに陥ることも多い．筆者は，ある程度データ数が多くて，母集団分布が正規分布にある程度近似的に従っていると考えられる場合は，パラメトリックな方法を用いられることをおすすめする．

多重比較を用いてはいけない場合

「そうか，たくさんのものを同時に比較する場合は，Studentのt検定を繰り返し用いてはいけないんだ．じゃあとにかくそんなときは多重比較を使用しよう」と思われれば，本稿はある程度成功したことになるが，ちょっと待って頂きたい．今少し正確に，多重比較を用いる必要性を強調してみたい．

「多数のデータの比較を同時に行い，有意差の存在を主張したい場合，Studentのt検定を繰り返し用いてはいけない．適切な多重比較法を用いるべきである．」

ここで，下線部に注目して頂きたい．生物工学や分子生物学の分野では，たいていの場合，何らかのデータの組み合わせに有意差があることを証明したいのだが，時には「有意差がない方がありがたい」場合も存在すると思われる．たとえば，いくつかの遺伝子が表現型に関係ないことを証明したい場合，その遺伝子を一つずつノックアウトして，表現型（目の大きさでもしっぽの長さでも何でもいいのだが）を測定，有意差があるかどうかを検定するとする．この場合，多重比較法を用いると，第2種の過誤の可能性が大きくなりすぎる．要は，「本当は有意差があるのに，誤って有意差が見つけられない可能性」が高くなるのである．多重比較は，あくまでも「本当は有意差のないものが，間違って偶然に有意差があると出てしまう確率」を制御するためのものであり，多数同時比較の場合，有意差検出をより厳しい目で見ていることになる．ということは，逆に「本当は有意差があるのに誤って有意差が見つけられない可能性」は，甘く見てしまわざるを得ない．

そもそも，統計的有意差検定は「有意差がある」ことを主張するのは得意だが，「有意差がない」ことを主張するのにはうまくできていない．極端なことを言えば，「AとBの母集団値は完全に等しい」ことを積極的に主張することは統計学以前の問題として不可能であることをご理解頂きたい．AとBが異なる母集団に属している限り，いかに両母集団が等質といえども，たとえ0.001%でも違いが存在するはずであり，そうなれば帰無仮説，

$H0：\mu_A=\mu_B$

が成立することは，正確には，ないのである．繰り返しになるが，統計的有意差検定は原理的にはあくまでも「有意差検定」であり，「同等性検定」でないことにご留意頂きたい．

その他

1）用いてはいけない多重比較法　多重比較法の中には，古くて，現在では使用されるべきではない手法が，

表2．代表的な多重比較法の分類

	パラメトリック	ノンパラメトリック
群間総当たり制の場合	Tukey 法	Steel-Dwass 法
対照群がある場合	Dunnett 法	Steel 法
群間に序列がある場合	Williams 法	Shirley-Williams 法

NII-Electronic Library Service

The Society for Bioscience and Bioengineering, Japan

高名なソフトの中に隠れ住んでいる場合がある．論文に「○△×□を用いました」と書けば，素人筋には信用されてしまうおそれもあるので結構怖い話ではあるのだが，残念ながら厳然と存在していて，しばしば論文中に散見されるのも事実である．極端な場合は，（さすがに希になったが）制約なしのLSD法が簡単に用いられていたりするし，Duncanの方法などは，時として見受けられる．これほど極端でなくても，制約ありのLSD法やNewman-Keuls法は，3群の比較の場合にのみ正しいのだが，これを4群の場合に単純に拡張して用いたりしているのは，そう希ではないので困ったものである．間違った方法をわざわざ解説するまでもないので省略するが，これらの方法は用いない方が無難である．制約付きLSD法を，FisherのLSD法などと呼ばれれば，名前負けして，つい使用してしまいそうだが，4群以上の比較に用いてはならない．たとえ高名なソフトがお墨付きを与えていようとも（もっとも，ソフトメーカーがお墨付きを与えているわけではないのだが，存在すると正しいものと思ってしまうのも人情だろう），間違った方法は用いてはならないのである．

2）FDR　繰り返しになるが，1回では起こりにくい現象も，何度も繰り返せば，1回くらいは起こっても不思議ではなくなるため，多数の数量の比較を同時に行い，かつ特定の比較の結果，片方の数量が有意に大きいなどと言うことを積極的に主張するときには，この「下手な鉄砲数打ちゃ当たる」現象に注意する必要があり，これが基本的な多重比較の考え方で，通常「FWER (family-wise error rate) を，公称の値～たとえば5%以下～に制御する」ことを目的に行われる．要は，「本当は有意でないものが，少なくともひとつ，有意とされてしまう確率」を小さくする方向に考えられている．しかしながら，この場合「本当は有意でないはずのものを棄却してしまう確率（FDR; false discovery rate）」もまた当然ながら自動的に考慮されているが，これを制御しているわけではない．したがって「少なくともひとつなどと言う厳しい条件でなくてもいいから，間違って棄却してしまう確率を制御したい」場合には不適である．

この，本当は有意でないはずのものを棄却してしまう確率は，以下のように定義できる．

$$\alpha_2 = \frac{F_1}{F_1 + T_2} \qquad (4)$$

残念ながら，この値を知ることは永遠にできない．しかし，その期待値を考えることは可能である．そこで，Benjamini and Hochberg[3]は，FWERでなくFDR，

$$E(\alpha_2) = E\left(\frac{F_1}{F_1 + T_2}\right) \qquad (5)$$

だけを制御する方法を発案した．紙数の関係で詳しくは述べないが，ポストゲノム時代，ゲノム解析やアレイデータの解析に，この手法は一定の有用性を持っていると思われる（比較の数が天文学的になるため）．このFDR制御で算出される危険率は，p-valueのかわりにq-valueと呼ばれる．近年はBenjamini-Hochberg法（BH法）に加え，その改良版ともいえるAdaptive BH法[4]，BY法[5]などが提唱されていて，FDR制御が主流になりつつあると言ってもいいかもしれない．

3）分散分析　今まで敢えて触れなかったのだが，多くの参考書には，多重比較法は1元配置分散分析法を行い，たとえば表1の場合なら，帰無仮説H0，

H0：$\mu_A = \mu_B = \mu_C = \mu_D = \mu_E$

H1：$\mu_i \neq \mu_j$ $\{\exists(i, j)\ ;\ i \neq j\}$

が棄却された場合にのみ用いると書かれている．ものは試しに先の表1のデータで分散分析を行ってみよう．一元配置分散分析はExcelで簡単に実行できる．結果は表4のようである．

「観測された分散比」はF分布に従うので，右端のF境界値（ここでは有意水準5%）に比べて十分に大きいため，帰無仮説H0は棄却され，対立仮説が採択される．さて，どことどこに有意差があるのか…は，通常の一元配置分散分析は教えてくれない．「どこかにある」とは教え

表3．FWERとFDR

	検定の結果	
	非有意	有意
有意差がないはずの比較	T_1	F_1
有意差があるはずの比較	F_2	T_2

表4．表1のデータの一元配置分散分析の結果

変動要因	変動	自由度	分散	観測された分散比	P-値	F境界値
グループ間	8.401	4	2.100	25.920	3.5E-11	2.578739
グループ内	3.647	45	0.081			
合計	12.048	49				

NII-Electronic Library Service

The Society for Bioscience and Bioengineering, Japan

てくれるが，そこまでである．あくまでも多重比較を目的とした場合だが，この種の分散分析は，筆者にはどうも cost effective だとは思えない．すべて手計算かせいぜい電卓に頼らざるを得なかった時代なら，多数のペアの比較を行うのは大変なので，一元配置分散分析で有意差が出たら行うことにしましょうという優しさがあったのだとも思うが，いいにつけ悪しきにつけ，すべてソフトウェアが行ってくれるような時代には，あまり意味がないように感じるのは筆者だけであろうか．そもそも多重比較と分散分析は原理的に性質の異なるものであり，絡めて考える必然性はないように思う．もちろん筆者は，分散分析が不必要だと言っているわけではないので念のため．現に筆者らが作成して試験的に公開している統計解析プログラムパッケージ，MEPHAS[6] にも，分散分析は含まれている．

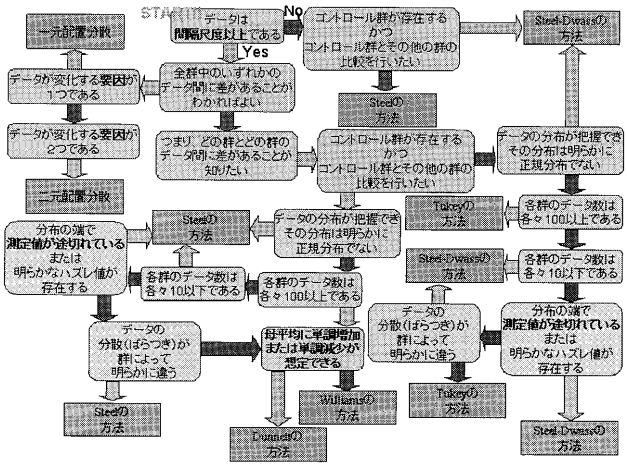
最後に

多重比較は，統計解析法の中でも最も厄介なものとして敬遠される向きがあるのは否めない．難しいというよりは複雑（平たく言えば「ややこしい」）のであるが，臨床現場では，人の命を左右しかねない問題でもあるので，疎かにはできない．残念ながら，比較的最近まで，分子生物学などの分野では，多重比較などまったく考慮に入れられていなかったり，誤った検定法（先述の Duncan 法など）が用いられていたりしたのも事実である．場合によっては論文の結論をひっくり返す場合も出てくるのだから，私たちは多少ややこしくても付き合わねばならない．本稿が，多重比較法を考える上で何かのきっかけになったのであれば幸いである．

最後の最後になってしまいましたが，本稿のきっかけを与えて頂きました大阪大谷大学薬学部，川瀬雅也教授に深く感謝致します．また，日頃より有益なディスカッションを行って頂いています大阪大谷大学薬学部，錦織理華助手，大阪大学大学院薬学研究科・微生物動態学分野／医薬情報解析学分野の岡本晃典助手をはじめ，研究室の皆さん，卒業生の皆さんに感謝致します．

文　献

［全般］永田　靖，吉田道弘（編）：統計的多重比較法の基礎，サイエンティスト社 (1997).
1) 樺　旦純：図説・数学トリック，王様文庫 (2003).
2) Takagi, T. *et al.*: *J. Comput. Aided Chem.*, **4**, 27 (2003).
3) Benjamini, Y. and Hochberg, Y.: *J. Roy. Statist. Soc. Ser. B*, **57**, 289 (1995).
4) Benjamini, Y. and Hochberg, Y.: *J. Edu. Behavioral Statist.*, **25**, 60 (2000).
5) Benjamini, Y. and Yekutieli, D.: *Ann. Statist.*, **29**, 1165 (2001).
6) 岡本晃典ら：http://www.gen-info.osaka-u.ac.jp/MEPHAS/



MEPHAS web画面より

NII-Electronic Library Service