〔日本生理誌（1996）**58**, 377-384〕

シリーズ「医学統計処理の問題点」

医学統計処理の問題点

—誤用のはなしⅣ—

栗谷典量
（久留米大学医学部小児科）

t検定の問題点 3

6．多群間の平均値比較（Comparing the Means of Several Groups）法の誤り

医学統計処理で誤用例の多いテーマとして，多群間平均値の比較法と，多重比較（Multiple Comparison）の不適切な適用の問題をあげねばなるまい．

2群間の平均値の有意差検定にはt検定が用いられ，3群以上（多群）の平均値の有意性検定には分散分析（analysis of variance：ANOVA）が用いられる．そこで，多群の検定を行う場合の対立仮説は“k個の母集団は等しくない”とするのであるから，これに対する帰無仮説が捨てられぬ限り，個々の群の差のt検定を行うのは問題があるわけである．多くの論文をみると，多群の実験結果の比較に当たり，多群中の2群の組み合わせ（ペア）を作り，例えばＡＢＣ3群の場合，Ａ：Ｂ，Ａ：Ｃ，Ｂ：Ｃ，の3通りの組み合わせ毎に，次々にt検定を繰り返す．この積み重ねをもって多群の全体を解析できたと錯覚する．この考え方は国内の医学関連雑誌では，一般的な手法として適用されているが，残念ながらこのやり方は基本的に欠陥解析で，正式に世界には通用しない方法なのである．実験，観察の対象が多群間比較の場合，結果の解析にはそれなりの知識と技術が必要であるが，国内の医学界においてはこの問題に関して，長年無視してきたというのが実情のようである．そのため，多群間の平均値の比較の適正処理に馴染まなかったようだ．多群間の比較の適性な処理とは，先ず最初に分散分析で“多群の平均値に差はない”とする帰無仮説が棄却出来てそこで初めて，個々の組み合わせの比較（多重比較）にかかれるのである．分散分析が，複数個の処理群全般について差があるかを問題とするのに対し，多重比較は，どの群とどの群に差があるのかを明らかにするために用いる統計的仮説検定の手法である．即ち多群平均の解析に当たり，第一段階として，多群全般を総合的に観察し（分散分析），そこで帰無仮説が棄却された（有意の）場合，第二段階として，次にどの群とどの群が異なるのかをはっきりさせたくなる．このために用いられる手法を総称して多重比較と呼んでいる．もし，第一段階の分散分析で仮説が棄却されない（有意でない）場合，第二段階の多重比較には手を付けない．多重比較の誤用パーンは，最初に行うべき解析，多群間の同時比較を手抜きする点が問題なのである．即ち，多群をまとめて比較することに対する関心が薄く，個々の群の平均値の比較に固執し，組み合わせ毎の2群の検定を繰り返すのが実情のようだ．

誤用例　6－1

うつ病の治療薬 fluvoxamine maleate のＬ群（25 mg），Ｈ群（50 mg），Ｉ群（イミプラミン）の3群の二重盲験試験．解析は3群の，投与期間・投与量（計量値）の比較であるが，ノンパラメトリック（U検定）で処理されている・解析手順は，先ず3群を2群の対3組に組み換え，つぎつぎにU検定を繰り返している．3群の同時比較（Ｈ検定）は行われてない．この論文には3群比較の項目が100種以上あるが，3群同時比較は全く念頭になく，すべて2群の対の組み合わせの検定のみを繰り返している．

表6最終全般改善度（追加解析2）は3群の全般改善度をTukey多重比較のみで終わっている．3群同時比較のＨ検定が手抜きされているので計算した結果は $H=3.3115$, $df=2$, $P=0.1999$ ns で3群の改善度に差はない．表6はＨ検定の結果の統計量を書き足せば，Tukeyの多重比較の値は削除できる．

表5． 投与期間，最高・最終1日投与量

項	目	L 群	H 群	I 群	U 検 定
投与期間（日）	～ 7	15	16	18	
	8 ～ 14	1	3	6	L：I p=0.219
	15 ～ 21	6	5	7	H：I p=0.431
	22 ～ 28	28	25	20	H：L p=0.690
	29 ～	5	5	5	
	平均±S. D.	20.6±10.6	19.5±11.1	18.0±10.3	
	不 明	3	3	0	
最高1日投与量（錠）	6	34	30	37	L：I p=0.701
	8	11	7	9	H：I p=0.168
	12	10	17	10	H：L p=0.298
	平均±S. D.	7.49±2.28	8.15±2.72	7.39±2.29	
	不 明	3	3	0	
最終1日投与量（錠）	6	35	30	37	L：I p=0.821
	8	10	7	9	H：I p=0.168
	12	10	17	10	H：L p=0.240
	平均±S. D.	7.45±2.29	8.15±2.72	7.39±2.29	
	不 明	3	3	0	

表6． 最終全般改善度（追加解析）（2）

薬 剤	著明改善	中等度改善	軽度改善	不 娠	やや悪化	悪 化	重篤に悪化	合 計	Tukey 多重比較検定	改善率（「中等度改善」以上）（%）
L群	14	9	9	2	3	0	0	37	N. S.	62.2
H群	9	8	7	6	1	4	0	35	N. S.	48.6
I群	8	13	6	4	5	2	0	38	N. S.	55.3

注）：検定結果は上から順にL群 vs H群，L群 vs I群，H群 vs I群で表示　　N. S.：not significant

この例のように多群の同時比較のことが，全く念頭にない解析法には次のような落し穴がある．多群の試験結果の検討に際し最初に行う処理は，多群全体をまとめて観察する姿勢である．例えば陸上競技400メートルの競争で，400メートルのタイムに関心示さず無視し，100メートル毎のラップタイムにのみ注意を払っていたのでは正しい結果の把握は望めない．また，野球のスコアで言うなら9回戦の合計得点を無視し，各イニングの得点だけ，あるいはヒット数・ホームラン数のみを見つめていても結果は同じである．どの選手が速く走ったのか，どちらのチームが勝ったのか，これでは正確な答えは得られない．400メートルのタイム，あるいは9回戦の合計得点を確認した後，次ぎにラップタイムや各イニングの得点についての検討を始めるのが順序だろう．現在，日本の医学雑誌で多群間の比較処理法について，厳格な処理を要求しているところは非常に少ない．即ち国内の医学関連では多重比較を誤用しても掲載拒否になることは，今日まであまり無かったのである．しかし，外国の医学雑誌（特に米・英）に投稿した場合，多群間比較，多重比較法について訂正の要求，あるいは掲載拒否の指摘を受けた例は

珍しくない．外国雑誌に投稿して，多重比較法の誤りを指摘された著者から，この指摘に対する対策手段の相談を受けることがあるが，国内誌からの指摘の話はあまり経験がない．ところが1995年秋，このテーマで初めての相談を経験した(西日本泌尿器科からの指摘)．

なぜｔ検定の繰り返しではいけないのか？

4群の比較試験を行った場合を例に考えてみよう．この群の各々のペアに対しそれぞれｔ検定を行うとすると，群(1：2)，群(1：3)，群(1：4)，群(2：3)，群(2：4)，群(3：4)の述べ6回のｔ検定を行う勘定になる．一般にｋ群の実験では ${}_kC_2=k(k-1)/2$ 回の検定の繰り返しになる．ところで検定の繰り返し回数が増加することの意義について考えてみる必要がある．同一母集団から得られたランダムサンプル群の平均値のペアを検定すると，検定する回数の増加につれて問題が生じてくる．有意差ありとは有意水準を0.05とした場合，20回に1回しか起こらないような，起こりにくいことが起きた，といって珍しい現象と判断するのである．では，同一母集団から10組のランダムサンプル群を作り，10群の平均値の比較をするとなると ${}_{10}C_2=45$ 回のｔ検定を行うことになる．45回の検定では単純に計算しても2回強の有意差をみる勘定になる．このことは，検定回数がふえるに従って有意になるチャンスが増加する(有意水準があまくなる)ということになる．そこで，その対策として検定回数に歩調を合わせて，有意水準を厳しくすることを考慮するのである．いま2つの検定を有意水準0.05で行った場合を例に考えてみる．この2つの検定で同時に統計学的に有意の場合，2つの有意水準は$0.05\times0.05=0.0025$となる．どちらの検定でも有意でない確率は$0.95\times0.95=0.9025$となる．2つの検定のうち少なくとも一方で有意になる確率は$1-0.9025=0.0975$である．したがって，2つの検定を用いてどちらか，または両方の平均のペアが等しくないと誤って判断する確率は，1回の検定での同じ過誤率のほぼ2倍になる(0.05：0.0975)．もし3群ならば，ペアの数は3になり，3つの検定すべてが有意でない確率は$0.95\times0.95\times0.95=0.8574$である．少なくとも1つで有意な確率は$1-0.8574=0.1426$となり14％を越えてしまう．始めは5％の有意水準で検定したつもりであったのだろうが，検定の繰り返しを考えもなくやっているうちに水準が甘く(有意になり易く)なっているのに気付かない．ここが問題なのである．有意差がで易くなる程度を次の式に示す．

少なくとも1つ，誤った検定結果を出す確率$=1-(1-\alpha)^n$

繰り返し回数ｎが増加すると，この確率は1回の検定の元にしている有意水準0.05より大きくなり，α(言い過ぎの危険率)は近似的にｎ倍になる．

群数	検定数	α 水準
2	1	0.0500
3	3	0.1426
4	6	0.2649
5	10	0.4013
6	15	0.5369
7	21	0.6594

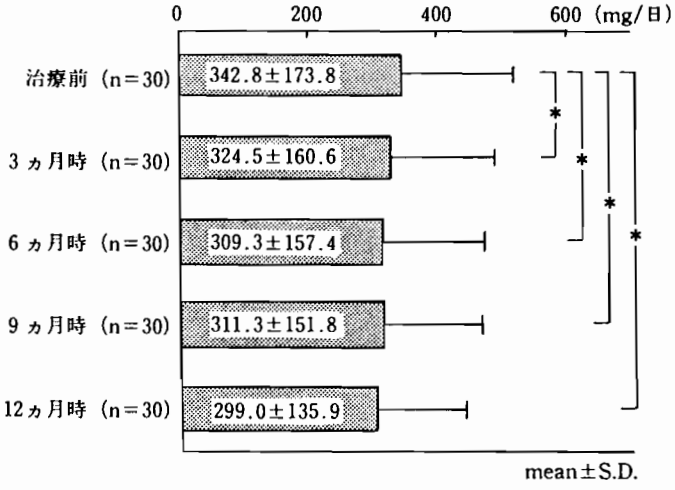
多重比較の問題は，対応のない多群間に限っての話ではなく，対応のある多群間(多時点間)の場合も全く同様に考えなければならない．対応のある多群の代表的データとして，時系列データが上げられる．時系列データとは時間の経過にしたがって定時毎に繰り返し測定して得られた一連の流れのデータ群のことである．

解熱剤投与後の体温の推移を，投与前，1時間後，2時間後，3時間後，合計4回測定した場合，投与後の体温の比較を，4回測定値の全組み合わせを考えれば6回，前値との比較だけなら，3回の検定の繰り返しになる．いきなり，6回または3回の対応のあるｔ検定のみですますと，多重比較処理の欠陥の対象になる．時系列の解析には，各時間の測定値とまとめて推移を検討する時系列分散分析の優先適用を考慮する．

誤用例　6－2

抗パーキンソン病薬の投与量の平均値の推移を検討している．

治療前3ケ月経過毎の平均値の paired t test のみで処理されているが，経時的データの時期に対する検討だから時系列分散分析を優先適用する．個々の時点との比較は，分散分析の結果次第でその次に．この試験の3ケ月毎の測定データに欠測値がないのは立派．



＊：$p<0.05$ (paired t-test)

注）L-Dopa 単味剤は L-Dopa/DCI 配合剤の L-Dopa 量に換算

図3．　各評価時期での L-Dopa 投与量の比較

誤用例　6－3

降圧剤の臨床効果．投与週毎の血圧の平均値の推移を観察しているが，解析は観察期と各週の Paired t test だけで終わっている．経時的データであるから，前誤用例6―2と同様，時系列分散分析を優先適用する．前例6―2の各時点の症例数にはバラツキがないが，例6―3の症例数は28～60で週毎のバラツキが大きい．

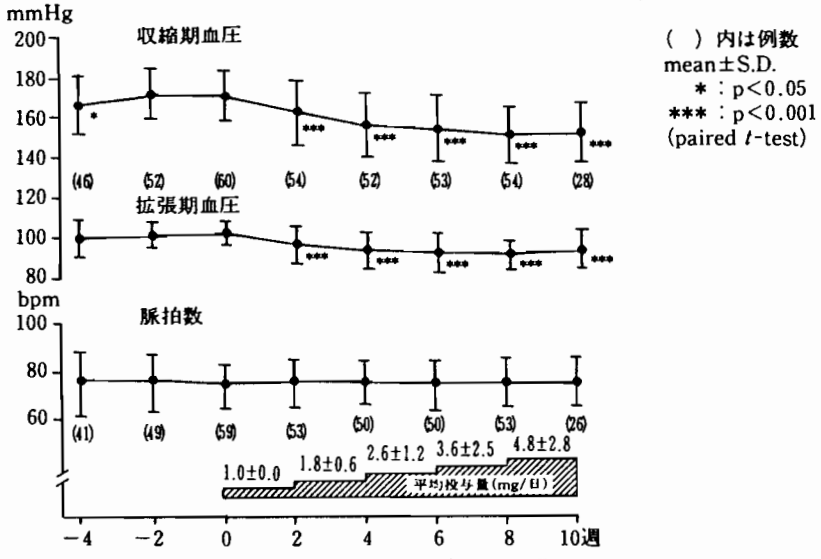


図3．　血圧・脈拍数の推移

多重比較の問題はｔ検定にまつわる認りだけでなく，順位検定，カイ二乗検定に関係する問題にも同様の誤りが数多く認められる．Mann-Whitney U test の場合を例にあげる．

例えば，治療効果の判定などで多用される「著効・有効・やや効・不変・悪化の５段階評価」を２群でU検定を行った場合，５段階評価を著効と有効以下の２段階評価に組み換え（この組み換え操作を尺度合わせと称する），あるいは有効以上とやや効以下の２段階に尺度合わせをする．こうして都合４回の検定の繰り返しが可能になる．このように初期の計画では５段階評価で判定されていたのを，解析に際し，２段階に組み換えて比率の比較検定を追加するやり方を常套手段とする著者は珍しくない．５段階評価を２段階評価に組み換えると４回検定が可能になるが，有意水準が５％→18.55％（約20％）と甘くなっている点はｔ検定の繰り返しと同様の誤用である．次ぎに尺度合わせの例をしめす．

	著効	有効	やや効	不変	悪化	合計	U test	有効以上	やや効以上
L 群	4	33	41	16	1	95		Fisher	Fisher
H 群	7	31	46	19	2	105	P=0.7345	P=0.7980	P=0.8425

L，H２群の治療効果に有意差は認められない．さらに，有効以上，やや有効以上，の有効率にも有意差は認められない．…この例の場合，U検定の結果，帰無仮説を棄却できないまま有効率の検定を始め，都合３回の検定の繰り返しを行っている点解析上の問題がある．

検定の繰り返しにより有意水準が甘くなる事に対する対策については，多数の研究がある．そのひとつ Ryan 法を紹介しておく．有意水準が甘くなる度合いは，近似的に検定の繰り返し回数倍ということから，単純に初期の有意水準を繰り返し回数で割った値を，新しい有意水準とするのである．もし平均値がｋ個あり，個々の組の平均について有意差を知りたいとき，検定の回数は ${}_kC_2$＝Ｌ回となるから，試験全体としての有意水準 α をＬで割った値まで下げておけば安全である．全体の水準に対し，Ｌで割った水準 α' を名義的（nominal）水準と呼んでいる．この対策法を Ryan 法という．比較群数が４群の場合を Ryan の法で確かめてみる．４群のときの検定回数は６回，有意水準を0.05とした場合，５％ Ryan nominal（α'）＝0.05/6＝0.00833 となる．Ryan nominal は検定回数で有意水準を単純に割っただけの値であり，理解は容易であるが，保守的である（厳くなる）．

誤用例　６－４

片頭痛に対するスマトリプタンの初期臨床Ⅱ相３用量（25 mg，50 mg，100 mg）試験，薬剤の投与デザインは corss over 法．改善度判定データは順序尺度である．３群比較にH検定を適用しているが，３群のデータには対応があるのでH検定は使えない．多重比較には，著名改善以上，軽度改善以上，で３回尺度合わせを行い，カイ二乗検定を行っているがこの検定も対応を無視しているので誤り．単なる誤りのみならず，計算値も誤り．この解析に於ける検定の繰り返しは，３群から２群の組み合わせを組む操作で３回，尺度合わせで３回，都合３×３＝９回の検定の繰り返しを行っている．その結果として，軽度改善以上で有意（P＜0.05）と書いてあるが，検算してみたが，有意でない．もっとも３群比較のH検定（H検定の使用は不適当だが使えるとしても）の結果が P＝0.4874 で多重比較を必要とするようなデータではない．仮に必要とした場合，S検定も使えない．当論文には60回以上検定が行われているが，正しい検定処理はなく全て不適当．この論文は cross over の解析法の学習未経験者の執筆と思われるが，順序効果の処理法の誤りの程度に至っては論外．cross over では時期効果がトラブルの原因となる事が定説であるが，時期効果の解析抜きでは薬効評価は不可能．特にこの試験デザインは時期が３期まであることからみて，時期効果，順序効果の解析法の研究，データの対応の有無と統計手法の使い方など解析法の基礎の学習を積む必要がある．３回 cross over 法の解析は無理と判断する．

表8. 頭痛改善度

投与量	症例数	著明改善	中等度改善	軽度改善	不変	悪化	H-検定	S-検定	x^2-検定		
									著明改善以上	中等度改善以上	軽度改善以上
25 mg	34	6(17.6) [17.6]	12(35.3) [52.9]	4(11.8) [64.7]	12(35.3)	0(0.0)	N. S.	N. S.	N. S.	N. S.	25<100* 25<50†
50 mg	32	6(18.8) [18.8]	11(34.4) [53.1]	10(31.3) [84.4]	5(15.6)	0(0.0)					
100 mg	32	7(21.9) [21.9]	12(37.5) [59.4]	9(28.1) [87.5]	4(12.5)	0(0.0)					

*：$p<0.05$，†：$p<0.10$

誤用例 6－5

抗パーキンソン薬，3投与量の臨床評価論文．データ尺度は順序尺度であるからH検定の適用は適性であるが，H検定の結果が有意でないのに，3薬剤の多重比較(x^2 検定)を行ったこと，多重比較の有意水準の調整を無視したのは欠点．

表7. 最終全般改善度

群	解析対象例数	著明改善	中等度改善	軽度改善	不変	悪化	判定不能	H検定[2)]	「中等度改善」以上	x^2 検定	95% 信頼区間
L群	36	1	12	17	3	0	3	p=0.3131	13 (36.1%)	L-M: p=0.6578	20.4% ～51.8%
M群	32	1	8	14	7	0	2		9 (28.1%)	L-H: p=0.2411	12.5% ～43.7%
H群	34	0	7	19	4	0	4		7 (20.6%)	M-H: p=0.6696	7.0% ～34.2%

1)：判定不能は除外した

誤用例 6－6

生後0～3月，4月～11月，1年以上の児，3群で牛脳自動酸化阻止能(AOA)，フェロキダーゼ比活性等をt検定のみで比較．分散分析が無視されている．

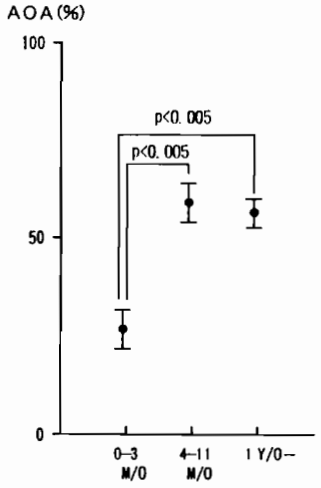


図1. 月齢別牛脳自動酸化阻止能，平均±標準誤差．
n=12(0～3M/O)，18(4～11M/O)，48(1Y/O～)

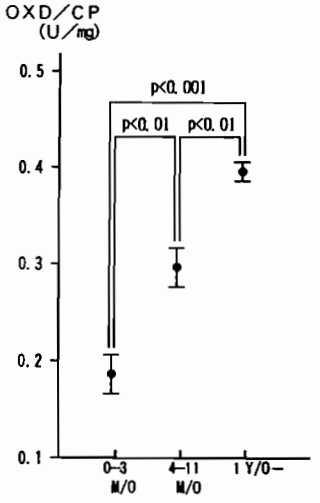


図3. 月齢別フェロキダーゼ比活性，平均±標準誤差．
n=12(0～3M/O)，18(4～11M/O)，48(1Y/O～)

誤用例　6－7

低出生体重児の生体電気インピーダンスを月齢別に5群でt検定で比較した論文5群から2群を取る組み合わせは10である．5群の ANOVA はない．

“＊”マークが5～8個付いているが，多重比較のため有意水準の調整を考慮して計算したところ，多くが消滅した．

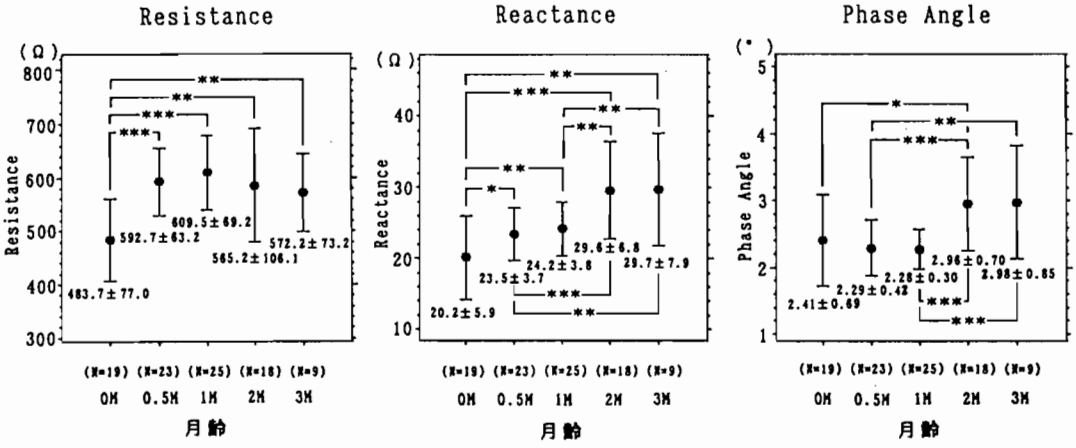


図3．　早期新生児以降の月齢での変化

＊：$p<0.05$，＊＊：$p<0.01$，＊＊＊：$p<0.001$

誤用例　6－8

この論文は前月の誤用例5―6で登場，皮膚科におけるビタミンD3の二重盲験論文である．実はこの試験デザインは同一患者の皮疹4カ所に4濃度の軟膏を塗布して比較するマッチドペアーである．つまり4カ所のデータは対応のあるデータである．従って最初に行う4群の比較に Krusal-Wallis H検定は使用出来ない(対応のないデータにしか適用出来ない検定法)．次ぎに行った多重比較には，Tukey 法が使われているが，Tukey 法は対応のあるデータには使用出来ない．症例数は41例が正しいが，124症例として登録されている．症例数の水増しは不正に当たる．症例数の不正，主目的の検定法の選択の誤りなど，試験計画に欠陥があるので部分訂正では修復不能．

	有用度								
薬剤群	極めて有用	有用	やや有用	どちらともいえない	有用でない	合計	有用率＊(%)	Kruskal-Wallis 検定	Tukey の多重比較検定
基剤	2 (4.9)	8 (19.5)	14 (34.1)	13 (31.7)	4 (9.8)	41	24.4	$p<0.01$	$p<0.01$，$p<0.01$，$p<0.01$
25 μg/g	10 (24.4)	15 (36.6)	11 (56.8)	5 (12.2)	0	41	61.0		
50 μg/g	16 (39.0)	17 (41.5)	5 (12.2)	3 (7.3)	0	41	80.5		
100 μg/g	18 (43.9)	17 (41.5)	4 (9.8)	1 (2.4)	1 (2.4)	41	85.4		

＊有用率：「有用」以上/解析症例数，( )：%

多重比較の方法論に関する研究は多数存在する．これらについては多くの推計学の著書に書かれているが，すべてをマスターして充分に使いこなせるまで習得することは，かなり困難である．しかし，多群を2群の順列組み合わせのペアーにして，何の考慮もなくt検定を繰り返すのは欠陥解析であることの理解は容易でろう．理解ついでに，

①多群の平均値の ANOVA で有意差が認められない場合，多重比較は行はない．
②多重比較では有意水準を，繰り返し回数に応じて厳しくする．
の条件を把握しておく事が重要である．
多重比較法の主なものを列挙しておく．

A．ボンフェロニー(Bonferroni)の不等式：
B．ライアン(Ryan)の法：群数が k の場合，有意水準を $\alpha' = \alpha / {}_kC_2$ に下げておく
C．テューキー(Tukey)の限界値 WSD 法：
D．シェーフィ(Scheffe)の法：
E．フィッシャー(Fisher)の LSD 法
F．ダネット(Dunnett)の法：
G．ダン(Dunn)の法：

多重比較法の方法論のコメントは略した，文献を一つ紹介しておく．

“医学統計学の活用” John C. Bailar Ⅲ, Fredrick Mosteller 編，津谷喜一郎・折笠秀樹監訳，サイエンティスト社，3,800円，
TEL 03-3253-8992

この翻訳書は統計の方法論の教科書ではない．P 169-185 が多群平均の比較の話題である．内容は論文の誤用例の分析・総括集である．

(説明の例　群間を比較するのに実行した解析は t 検定の繰り返しだけだった)

多重比較について，後書き

これまで数多く，多群の比較検定の処理について相談を受けた．多群の処理を実行して，依頼者に渡す演算結果には，必ず多群の同時比較検定の結果を冒頭に記載するようにしている．ところが著者が，その肝心の計算結果を論文の原稿にする時点で，しばしば無視，あるいは切り捨てられてしまうことがある．同時比較を行った際に算出した分散分析や，Kruskal-Wallis の H 検定の結果などの統計量が，いつの間にか消滅し，原稿のどこにも見当たらなくなっているのである．「ANOVA はどこへ行ったのでしょうか？」と聞いてみると「あれは重要と思われなかったから外しました」．又は「何か数字があったが，よくわからなかったので無視した」と，著者達は多群同時比較の意味と理由を理解しないまま，処分しているようである．例えば，3 群比較の場合一般的に著者達は，個々の 2 群の組み合わせの 3 種の t 検定には，強い関心を示すが，3 群の同時比較 ANOVA には注意を払いたがらないのが実情のようである．

さらに，ANOVA の結果が有意であろうとなかろうと，結果の如何にかかわらず，多重比較の統計量はすべて欲しいと思う人が多い．ANOVA で有意にならなかったため，多重比較をやらないままにしておくと，どうしても 2 群毎の比較をして “＊” マークを入れたいと希望する人がある．

医学者は，推測統計学・確率論・解析技術の専門家として特に養成されている訳ではないが，知識として要請される立場にあることは認識しておきたい．