GÖTEBORG UNIVERSITY

Department of Statistics

RESEARCH REPORT 1994:5
ISSN 0349-8034

COMPARING POWER AND MULTIPLE SIGNIFICANCE LEVEL FOR STEP UP AND STEP DOWN MULTIPLE TEST PROCEDURES FOR CORRELATED ESTIMATES

by

Bo Palaszewski

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S-411 25 Göteborg
Sweden

COMPARING POWER AND MULTIPLE SIGNIFICANCE LEVEL FOR STEP UP AND STEP DOWN MULTIPLE TEST PROCEDURES FOR CORRELATED ESTIMATES

Bo Palaszewski

Department of Statistics, Viktoriagatan 13, S-41125 Göteborg, Sweden. E-mail: Bo.Palaszewski@statistics.gu.se

KEY WORDS; Multiple level of significance, step up test procedure, step down test procedure, power comparisons, product correlation structure, finite sample sizes, Monte Carlo simulation.

ABSTRACT

We consider hypothesis testing problems arising in e.g. the context of comparing k treatments with a control. The case of equi-correlated estimates is mainly discussed, although also unequal correlated estimates (e.g. unequal sample sizes for the treatments, when compared to a control treatment) are mentioned briefly. So called step down test procedures are compared with step up test procedures, with respect to power functions. Comparisons of rejected null hypotheses are also performed. No practical differences in performances between step up and step down test procedures could be found for finite sample sizes.

1. INTRODUCTION

Let's assume a standard normal theory linear model setting. Consider parameters $\theta_1,\theta_2, \ldots ,\theta_k$, $k\geq 2$. Further, let $\hat{\theta}_1,\hat{\theta}_2, \ldots ,\hat{\theta}_k$ be unbiased least squares estimates of the parameters $\theta_1,\theta_2, \ldots ,\theta_k$. They are assumed to be jointly normally distributed with variance $\text{var}(\hat{\theta}_i)=\sigma^2\tau_i^2$ and correlation $\text{corr}(\hat{\theta}_i,\hat{\theta}_j)=\rho_{ij}$, for i,j=1, ... ,k. Also, let the correlation coefficients have a *product* form $\rho_{ij}=\lambda_i\lambda_j$. Further we have $\{\tau_i^2\}$ for i=1, ... ,k, and $\{\rho_{ij}\}$ for

i,j=1, ... ,k, are known constants defined by the design. σ^2 is the unknown error variance. Let S^2 be an unbiased estimate of σ^2 with ν df. Then $\nu S^2/\sigma^2$ is distributed as a χ^2_ν variate, independent of $\hat{\theta}_i$. Finally let $\underline{\theta}$ be the vector $(\theta_1,\theta_2, \ldots ,\theta_k)$.

One example of this setting is comparison of k treatment means with a control mean in a one-way layout with n_0 observations in the control group and n_i observations for each of the treatment groups, where the groups are independent. Then

$$\lambda_i = 1\Big/\sqrt{1+\frac{n_0}{n_i}}$$

and $\tau_i^2 = 1/n_i + 1/n_0$, i=1, ... ,k. Mostly we are concerned with the equi-correlated case, where all estimates are correlated with a common correlation coefficient. In this example, clearly this is equivalent to the balanced design with $n_i=n$, for i=1, ... ,k, which also gives that $\rho=\lambda^2$. The unbalanced design is then identical to the case when the correlation between two estimates, $\text{corr}(\hat{\theta}_i,\hat{\theta}_j)=\rho_{ij}$, are unequal (i.e. $\rho_{ij}\neq\rho$).

The parameters of interest are $\theta_i=\mu_i-\mu_0$. The hypotheses are $H_i:\theta_i=0$ vs. A_i: $\theta_i>0$. The test statistics used are $(t_1,t_2, \ldots ,t_k)$, where $t_i=\hat{\theta}_i/s\tau$, i=1, ... ,k. This set of test statistics are used in the two stepwise test procedures discussed in this paper. $H_1,H_2, \ldots ,H_k$ are labelled so that the statistics t_i are ordered by increasing value $t_1<t_2< \ldots <t_k$. The multiple test procedure's critical constants satisfy the monotonicity condition $c_1<c_2< \ldots <c_k<\infty$, where c_i is the critical constant to be used with t_i.

The step down test procedure starts by testing if any hypotheses could be rejected. If t_k is sufficiently large, H_k is rejected, and the procedure continues by testing if H_{k-1} and so on. If any hypothesis is not rejected, then all of the remaining hypotheses including this hypothesis are accepted. The step up procedure starts by testing the hypothesis corresponding to t_1, H_1, the least significant test statistic. The procedure continues by testing H_2 only if H_1 was not rejected, and so on. The procedure stops when a hypothesis was rejected. This and all not yet tested hypotheses are then rejected.

2. STEP UP AND STEP DOWN MULTIPLE TEST PROCEDURES

Multiple test procedures are procedures which take into account for possible *multiplicity effects*. Such effects could result from applying many tests to the same data material. If all tests are performed by applying each test separately at significance level α, the resulting risk to reject one or more true hypotheses could well exceed α. Multiple test procedures are aiming to control the multiple significance level at some prechoosen level. The multiple level of signicance is defined as the risk to reject one or more true hypotheses, which ever and how many they are.

Dunnet (1955) suggested a single step multiple test procedure for the case of comparing k treatments with a control treatment. Equi-correlated data was assumed. Later this procedure was refined by Marcus et al (1976), by using the concept of closed test procedures (Gabriel, 1969). They showed that the proposed test procedure was more powerful than the single step version. This general procedure was shown to be equivalent to the step down test procedure by Holm (1977). Naik (1975) proposed a step down test procedure for the special case of k comparisons with a control treatment. This problem could also be regarded as a selection problem (see e.g. Gupta and Sobel 1958). This is not treated further in this paper.

The step down multiple test procedure starts by testing the hypothesis corresponding to the most significant test statistic and continues by testing the hypothesis corresponding to the next most significant test statistic. The procedure stops the first time a hypothesis is not rejected. All the previously hypotheses are then rejected. More specifically the step down procedure is as follows: Order the test statistics, $t_{(1)}<t_{(2)}< \ldots <t_{(k)}$, and its corresponding hypotheses $H_{(1)},H_{(2)}, \ldots ,H_{(k)}$. Reject any $H_{(i)}$ iff $H_{(j)}$ is rejected for $j=k, \ldots ,i+1$ and $t_{(i)}\geq c_i$. This procedure controls the multiple level at a prechosen level α.

The upper α point of the distribution of max T_i, the maximum of $T_1,T_2, \ldots ,T_m$, which have a m-variate t-distribution with ν degrees of freedom and a common correlation coefficient ρ, for $m=1, \ldots ,k$, are used to determine the critical test constants $c'_m=t^{(\alpha)}_{m,\nu,\rho}$ for the step down test procedure. The corresponding two-sided procedure

exists. Bechhofer and Dunnet (1988) has published tables with critical constants for different m, α, ν and ρ.

These tables are for the case of balanced designs, i.e. a common ρ for all test statistics. For the unbalanced case with arbitrary ρ_{ij}'s, a computer program (see Dunnet, 1985 and Dunnet & Tamhane 1991) is necessary to obtain critical values, because of the large number of different possible configurations with ρ_{ij}'s.

The step up multiple test procedure, proposed in Dunnet and Tamhane (1992), goes as follows: Start to test the hypothesis corresponding to the the least significant test statistic, t_1. If this hypothesis, H_1, is not rejected proceed to test hypothesis corresponding to the the next least significant test statistic, t_2. The procedure stops the first time a H_i is rejected. This H_i and all remaining H_j is then rejected. This procedure controls the multiple significance level at a prechosen α-level.

The critical constants for the step up procedure, developed by Dunnet & Tamhane (1991), is harder to determine than for the step down procedure, where the critical constants for different values of m can be computed independently of each other. Critical constants for the step-up procedure are determined by solving the following equation recursively for c_m given $c_1, \ldots, c_{m-1}$:

$$P[(T_1,T_2, \ldots ,T_m)<(c_1,c_2, \ldots ,c_m)]=1-\alpha$$

for m=1, ... ,k. $T_1,T_2, \ldots ,T_m$ have a central m-variate t-distribution with ν df and correlation matrix $\mathbf{R_m}$, which is the correlation matrix corresponding to the m smallest t test statistics. $(T_1,T_2, \ldots ,T_m)<(c_1,c_2, \ldots ,c_m)$ denotes that $T_{(1)}<c_{(1)}$, $T_{(2)}<c_{(2)}$, ... ,$T_{(m)}<c_{(m)}$, where $T_{(i)}$ and $c_{(i)}$ are the ordered T_i and c_i. These c_i's are also assumed to satisfy the monotonicity condition, although it has not been possible to show analytically for m>2. Dunnet & Tamhane (1992) conjectures that the condition is satisfied, and support the conjecture by numerical computations for $m \leq 8$ and $\alpha=0.05$. Values of critical constants are tabulated in Dunnet & Tamhane (1992). It is comparatively easy to calculate these c_i's if the correlation between all t_i and t_j are equal. If not, the computing of the c_i would also depend on the observed ordering among the test statistics. This requires the

solution of an equation involving an integral in multiple dimensions, which requires a very difficult numerical integration. These computations are required to repeat for each problem. Dunnet & Tamhane (1994) proposes two different approximate solutions to the case when $\rho_{ij} \neq \rho$. One is to replace the unequal correlation coefficients by their arithmetic averages, and then using the tables for the equal correlation case. Conjectures are made that this type of approximation results in conservative critical limits, if the assumption about product structure for the ρ_{ij}'s, are true. This has not yet been been proven by analytical results. Some calculations made by the author indicates that the approximate critical limits estimates the true limits with a surprisingly good precision, despite designs which are quit unbalanced.

3. COMPARING MULTIPLE TEST METHODS

In general, it is a difficult problem to compare different multiple test procedures. There exists no satisfying general concept concerning optimality of multiple test procedures (see Finner 1994). Several methods have been proposed for comparison of multiple tests, but unfortunately, the theoretical results are sparse. For the most cases, it is nearly impossible to obtain theoretical results concerning the power of such complicated test situations, as that of multiple testing. A consequence is that most power comparisons of multiple test procedures are carried out by Monte Carlo simulation studies. One exception is the result of Spjøtvoll (1972).

Tests for a single hypothesis are often compared in terms of their power functions. One often used definition of multiple power is the probability to reject a certain subset of false hypotheses, for multiple tests with given multiple significance level α. This is in line of the definition of the *P-subset power*, of Einot and Gabriel (1975). The P-subset power definition focuses on rejecting some on forehand selected false hypotheses. If P denotes a specific hypotheses, then the power is to be interpreted as the probability to reject that particular hypotheses. Another applicable notion of multiple power is the *overall power*. This is defined as the probability to reject all false hypotheses. This notion was used by Welsch (1977). In line of the proposition of Ramsey (1978), we could also define the

power as the *probability to reject at least one false hypothesis.* This could be any hypothesis. Generalisation to subsets are obvious.

Suggestions have been made to compare multiple tests pointwise and simultaneously in all components (Finner 1994). He uses the following definitions of power: An multiple α-level test is *not less* (in power to reject hypotheses, true or not) than another α-level test, if the power to reject a given subset of hypotheses are at least as large as the other test, for all possible values of the test procedures. One test procedure is *greater than* the other if its power is greater than or equal the second test procedure at most values, and greater than for at least one value. These measures are then used for obtaining results about admissiblity of multiple test procedures. Even when using a trivial loss function the results are scarce and very limited. Admissibility is only possible to prove for some very specialized cases.

Another important criterion when comparing multiple test procedures, is to study the (expected) number of rejected hypotheses. This measure was advocated for by Spjøtvoll (1972), in the case of finite families. With a family is here meant *any collection of inferences for which it is meaningful to take into account some combined measure of errors.* He suggested that the (expected) number of rejected null hypotheses was the error level that should be controlled for finite families of hypotheses, not the multiple significance level as defined above in this report. The reason was amongst other that this later definition imposes a penalty in direct proportion to the number of errors, while the multiple significance level definition corresponds to a zero-one loss function. He gave the following example '*Suppose a statistician uses (the expected number of false rejections)* γ *=0.05, then in average for every twentieth problem he makes one false statement. On the other hand if he uses (the probability to falsely reject one or more hypotheses)* α*=0.05, then in average for every twentieth problem he makes false rejections, but he does not know how many false rejections he makes*'. His definition of error level also gives an upper bound on the multiple level of significance, so controlling the former also controls the latter. It could be argued which error level that should be used when applying a multiple test procedure, but certainly it is wise to use the definition of Spjøtvoll, when comparing procedures. This is in contrast to the often used error level as defined by the multiple level of significance. The former error level is an upper bound

on the latter one. It seems to be necessary to study the expected number of rejected null hypotheses when examining the power of different methods, since there could be differences with respect to the number of rejected null hypotheses, as well as the power of the methods.

With this constraint he showed that amongst other results, the single step test procedure for comparing k groups parameter value with a control groups parameter value, as proposed by Dunnet (1955), was optimal in the sense to maximize the minimum (average) power over specified subsets of parameter spaces. The proposed optimality measures are aiming at to maximizing the performances of the individual tests. It does not tell us anything about the probability of rejecting several false hypotheses, i.e. nothing about their simultaneous performance.

4. SIMULATION RESULTS

To investigate some different aspects of the step-down (SU) and step-up (SD) procedures, simulation studies were performed, for which some chosen results are given here. First, let Z_i, for i=0,1, ... ,k, be independent N(0,1) random variables. Further, let U be a $\sqrt{\chi^2_\nu/\nu}$ random variable, with ν df, independent of Z_i. The test statistics could then be written as

$$T_i = \frac{\left\{\sqrt{1-\rho}\,Z_i - \sqrt{\rho}\,Z_0\right\}}{U}$$

for $1 \le i \le m$, $m \le k$. Further, for power simulation studies, we use

$$T_i = \frac{\left\{\sqrt{1-\rho}\,Z_i - \sqrt{\rho}\,Z_0 + \sqrt{\rho}\,D_i\right\}}{U}$$

to obtain test statistics for false hypotheses, where D_i is a shift parameter giving the difference from groups with true hypotheses. All simulation results presented in this report were obtained from the following set up: Normal pseudo-random variates Z_i,

$Z_i \sim N(0,1)$, were generated. U was generated with $\nu=27$ df, independent of Z_i, i=0, ... ,k. The common correlation coefficient ρ was taken to be 0.5. The pseudo-random deviates were generated for some different configurations. Each configuration was replicated 1000 times.

To estimate the observed multiple α-level, counts were made the first time in each replicate a true hypothesis was rejected at level α. This is in accordance to the definition of multiple significance level, as put in Holm (1977). To estimate the power of a method, counts were made when the procedure succeeded in rejecting all false groups. Counts was also made for the first rejected false hypothesis. This was done for different values for a suitable set of shift parameters.

We first consider the definition of power that require all false hypotheses to be rejected. In Dunnet and Tamhane (1992), it was conjectured that if exactly one hypothesis is false, then SU is uniformly more powerful than SD, for all values of the shift parameter. Further, they found that SU dominates SD uniformly when all hypotheses are false. The numerical results given in their paper, also gave that when an intermediate number of hypotheses are false, SU is more powerful than SD for small departures from the null values, while SD is more powerful in the other case. The other definition of power used in Dunnet and Tamhane (1992), i.e. the probability to reject at least on false hypothesis, resulted in the following result: Again SU dominates SD uniformly when all hypotheses are false. Also that the advantage of SU increased with increasing k, particularly at low levels of power. It was also found that when exactly one hypothesis is false, the advantage of SD decreases with k and is in most cases negligibly small. Dunnet & Tamhane (1992) presented the result *that the SU procedure had a non negligible power advantage only in those situations where most hypotheses are false and it is desired to reject all of them.* They also concluded that this power advantage increases with the number of false hypotheses. They also stated that even for the case of only a few false hypotheses, the SU procedure was only marginally worse than SD procedure to reject them. But their result was calculated with the degrees of freedom assumed to be large, i.e. $\nu=\infty$, and for number of groups k=6 and lower. For more realistic assumption about the degrees of freedom, any practical advantage for the SU procedure as compared to the SD procedure, seems to vanish. The results published

here, with k=8 and ν=27 shows that the observed level of significance were within the level aimed at, i.e. lower than or equal with 5%, for both SU and SD. We obtained results pointing towards the conclusion that the power differences of any practical magnitude, in favour of the SU procedure, was negligible for any configuration of true and false hypotheses. Simulation results supporting this, are given for the case that 8, 4, and 1 hypotheses are false, when a group size of 8 was considered. Other results not published in this report, shows that although the results is in accordance with the results in Dunnet & Tamhane (1992), the differences are not significant (at 5%), not even for ν=72.

Table 1-8 were obtained for shift parameters D_1=0.5, D_2=1.0, D_3=2.0, D_4=4.0, D_5=8.0. The same value on the shift parameter was applied to all groups in the set of groups with $\mu_i > \mu_0$, for all $i \notin I$, where I is the set of true hypotheses. Table 1-2 displays the result for the case of 8 false hypotheses out of a total of 8 hypotheses. No consistent pattern were found. The power differences were negligible for both of the power definitions, i.e. the probability to reject all false hypotheses and the power to reject at least one arbitrary false hypothesis.

Table 1 Probability to reject 8 false for the configuration with 8 false hypotheses.

D	SU	SD	i
0.5	0.00190	0.00190	1
1.0	0.00610	0.00620	2
2.0	0.05560	0.05650	3
4.0	0.51220	0.51200	4
8.0	0.99970	0.99960	5

Table 3 displays the result for the case of 1 false hypothesis out of a total of 8 hypotheses. Again, no consistent pattern were found. In Table 4, the observed significance level are given for the same configuration. Both procedures are within the multiple level of significance aimed at, 5%. The increase in observed significance level with increasing **D** is mainly due to the fact that for high values of **D**, we certainly rejects all the false hypotheses in their correct positions before the procedure stops, and are thence more exposed to the risk to reject one or more true hypotheses. The results given

pointing towards the conclusion that the power differences of any practical magnitude, in favour of the SU procedure, was negligible for any configuration of true and false hypotheses. Simulation results supporting this, are given for the case that 8, 4, and 1 hypotheses are false, when a group size of 8 was considered. Other results not published in this report, shows that although the results is in accordance with the results in Dunnet & Tamhane (1992), the differences are not significant (at 5%), not even for ν=72.

Table 1-8 were obtained for shift parameters D_1=0.5, D_2=1.0, D_3=2.0, D_4=4.0, D_5=8.0. The same value on the shift parameter was applied to all groups in the set of groups with $\mu_i > \mu_0$, for all $i \notin I$, where I is the set of true hypotheses. Table 1-2 displays the result for the case of 8 false hypotheses out of a total of 8 hypotheses. No consistent pattern were found. The power differences were negligible for both of the power definitions, i.e. the probability to reject all false hypotheses and the power to reject at least one arbitrary false hypothesis.

Table 1 Probability to reject 8 false for the configuration with 8 false hypotheses.

D	SU	SD	i
0.5	0.00190	0.00190	1
1.0	0.00610	0.00620	2
2.0	0.05560	0.05650	3
4.0	0.51220	0.51200	4
8.0	0.99970	0.99960	5

Table 3 displays the result for the case of 1 false hypothesis out of a total of 8 hypotheses. Again, no consistent pattern were found. In Table 4, the observed significance level are given for the same configuration. Both procedures are within the multiple level of significance aimed at, 5%. The increase in observed significance level with increasing **D** is mainly due to the fact that for high values of **D**, we certainly rejects all the false hypotheses in their correct positions before the procedure stops, and are thence more exposed to the risk to reject one or more true hypotheses. The results given

above were also exhibited for the configuration with 4 false hypotheses out of 8 hypotheses. Other investigated configurations resulted in about the same result.

We also investigated the distribution of rejected true H_i, i.e. how many times exactly one true hypotheses are rejected, exactly two true hypotheses are rejected, and so on. The distribution of rejected true hypotheses are given in Table 8, for the configuration with 4 false hypotheses out of 8 hypotheses. There is no differences of any practical magnitude between the two procedures. This frequency distribution could be used to form the expected number of true rejected hypotheses. The results gives that there is no detectable difference between the two procedures, with respect to the frequency distribution of wrongly rejected hypotheses. This pattern was also found for other configurations of true and false hypotheses.

Table 2 Probability to reject at least one false hypotheses. Same configuration as in Table 1.

D	SU	SD	i
0.5	0.10640	0.10600	1
1.0	0.19340	0.19340	2
2.0	0.45500	0.45310	3
4.0	0.92640	0.92640	4
8.0	1.00000	1.00000	5

Table 3 Probability to reject one false hypothesis, for the case that 1 is false out of 8 hypotheses.

D	SU	SD	i
0.5	0.02210	0.02200	1
1.0	0.04830	0.04820	2
2.0	0.16200	0.16200	3
4.0	0.61380	0.61400	4
8.0	0.99820	0.99820	5

Table 4 Observed multiple significance level with 7 true and 1 false hypotheses.

D	SU	SD	i
0.5	0.04630	0.04620	1
1.0	0.04880	0.04680	2
2.0	0.04820	0.04850	3
4.0	0.04940	0.04980	4
8.0	0.04970	0.05010	5

Table 5 Probability to reject all 4 false hypotheses out of 8 hypotheses.

D	SU	SD	i
0.5	0.00230	0.00200	1
1.0	0.00640	0.00630	2
2.0	0.04110	0.04040	3
4.0	0.39640	0.39700	4
8.0	0.99670	0.99670	5

Table 6 Probability to reject at least one false hypotheses out of 4 false hypotheses when there is 8 groups.

D	SU	SD	i
0.5	0.06940	0.06940	1
1.0	0.13120	0.13120	2
2.0	0.34890	0.34820	3
4.0	0.86410	0.86290	4
8.0	1.00000	1.00000	5

Table 7 Observed multiple significance level, same configuration as in Table 5-6.

D	SU	SD	i
0.5	0.03500	0.03480	1
1.0	0.03770	0.03730	2
2.0	0.04450	0.04370	3
4.0	0.05200	0.05240	4
8.0	0.05300	0.05330	5

5. CONCLUSIONS

The conclusions from this report gives no support for the conclusion that SU procedures are preferred before SD procedures, when analysing data with the given correlation structure, if the analysis was performed on situations with small number of degrees of freedom. Support for this conclusion comes both from the power studies as well as the detailed study of structures of wrongly rejected hypotheses. Both aspects should be taken in to account when comparing multiple test methods, since the power only gives information about the method's ability to reject false hypotheses. The most often used definition of multiple level of significance, the probability to reject one or more true hypotheses, disregarding which they are, does not take into account that a penalty should be given a method that more often rejects more than one true hypothesis. The increased power, as compared to some other method, could then emanate from an increase in the number of rejected true hypotheses. This is not the case here since the power seems to be equal for the two methods compared, as well as the structure of rejected true hypotheses. This makes the use of the SU method, with complex computations to determine critical test constants, less motivated to use in applied situations. Hence, it must be more compelling to use the more easily applied SD method. This conclusion does not preclude that further research might show that SU procedures are useful.

Table 8 The frequency distribution of rejected true hypotheses.

	Number	SU %	SD %
	1	0.02570	0.02570
D=0.5	2	0.00650	0.00670
	3	0.00160	0.00120
	4	0.00120	0.00120
	1	0.02700	0.02670
D=1.0	2	0.00720	0.00740
	3	0.00210	0.00180
	4	0.00140	0.00140
	1	0.03120	0.03080
D=2.0	2	0.00860	0.00850
	3	0.00300	0.00270
	4	0.00170	0.00170
	1	0.03700	0.03730
D=4.0	2	0.00920	0.00970
	3	0.00400	0.00350
	4	0.00180	0.00190
	1	0.03800	0.03820
D=8.0	2	0.00920	0.00970
	3	0.00400	0.00350
	4	0.00180	0.00190

ACKNOWLEDGEMENTS

This work has been supported by the Swedish Council for Research in the Humanities and Social Sciences.

BIBLIOGRAPHY

Bechhofer, R.E. and Dunnet, C.W. (1988), Tables of percentage points of multivariate t distributions, in *Selected Tables in Mathematical Statistics*, **11**, American Mathematical Society, Providence, Rhode Island.

Dunnet, C.W. (1955), *A Multiple Comparison Procedure for Comparing Several Treatments With a Control*, Journal of American Statistical Association, **75**, 1096-1121.

Dunnet, C.W. (1985), *Multiple Comparisons Between Several Treatments and a Specified Treatment*, Lecture Notes in Statistics, **35**, Linear Statistical Inference, eds. T. Calinski and W. Kloneski, Springer Verlag, 39-46.

Dunnet, C.W. and Tamhane, A.C. (1991), *A Step-Down Multiple Test for Comparing Treatments with a Control in Unbalanced One-way Layouts*, Statistics in Medicine, **10**, 939-947.

Dunnet, C.W. and Tamhane, A.C. (1992), *A Step-Up Multiple Test Procedure*, Journal of the American Statistical Association, **87**, 162-170.

Dunnet, C.W. and Tamhane, A.C. (1994), *Step-Up Multiple Testing of Parameters With Unequally Correlated Estimates,* to be published.

Einot, I. and Gabriel K.R. (1975), *A Study of the Powers of Several Methods in Multiple Comparisons*, Journal of American Statistical Association, **70**, 574-583.

Finner, H. (1994), *Testing Multiple Hypotheses: General Theory, Specific Problems, and Relations to Other Multiple Decision Procedures*, Habilitationsschrift, Trier.

Gabriel, K.R. (1969), *Simultaneous test procedures - some theory of multiple comparisons*, Annals of Mathematical Statistics, **40**, 224-250.

Gupta, S. S. and Sobel, M. (1958), *On selecting a subset which contains all populations better than a standard*, Annals of Mathematical Statistics, **29**, 235-244.

Hochberg, Y. and Tamhane, A.C. (1987), *Multiple Comparison Procedures*, Wiley, New York.

Holm, S.A. (1977), *Sequentially Rejective Multiple Test Procedures*, Statistical Research Report 1977-1, University of Umeå, Sweden.

Marcus, R. Peritz, E and Gabriel, K.R. (1976), *On Closed Test Procedures With Special Reference to Ordered Analysis of Variance*, Biometrika, **63**, 655-660.

Naik, U. D. (1975), *Some Selection Rules for Comparing p processes with a standard.* Communications in Statistics, **4**, 519-535.

Ramsey, P.H. (1978), *Power Differences Between Pairwise Multiple Comparisons*, Journal of American Statistical Association, **73**, 479-487.

Spjøtvoll, E. (1972), *On the optimality of some multiple comparison procedures*, Annals of Mathematical Statistics, **43**, 398-411.

Welsch, R.E. (1977), *Stepwise Multiple Comparison Procedures*, Journal of American Statistical Association, **72**, 566-575.

1993:1	Frisén, M & Åkermo, G.	Comparison between two methods of surveillance: exponentially weighted moving average vs cusum
1993:2	Jonsson, R.	Exact properties of McNemar's test in small samples.
1993:3	Gellerstedt, M.	Resampling procedures in linear models.
1994:1	Frisén, M.	Statistical surveillance of business cycles.
1994:2	Frisén, M.	Characterization of methods for surveillance by optimality.
1994:3	Frisén, M. & Cassel, C.	Visual evaluation of statistical surveillance.
1994:4	Ekman, C.	A comparison of two designs for estimating a second order surface with a known maximum.