

SIMPLE STEPWISE TESTS OF HYPOTHESES AND MULTIPLE COMPARISONS

EUGENE SENETA AND JOHN T. CHEN

School of Mathematics and Statistics

University of Sydney

N.S.W. 2006. AUSTRALIA.

Department of Mathematics and Statistics

Bowling Green State University

Bowling Green OH 43402

U.S.A.

ABSTRACT. Holm's (1979) step-down and Hochberg's (1988) step-up procedures for tests of multiple hypotheses are simple to apply and are widely used. Holm's procedure controls the familywise error rate (FWE), while Hochberg's is more powerful. This paper investigates a step-down procedure (labelled CS) of Seneta and Chen (1997) which is a sharpening of Holm's, takes into account the degree of association between test statistics, and also controls the FWE. Computation for the CS procedure may be minimized by using the procedure as an adjustment to Holm's. The computational steps are detailed, and the adjustment then illustrated by an application to a text-book example of multiple comparisons, in which step-wise procedures are shown to perform better than the usual Tukey T-comparison. Simulation investigations in a standard comparison with a control setting show that the CS step-down procedure is more powerful than Hochberg's step-up procedure and the procedure of Simes (1986), especially in regard to error rate, and not much less powerful than an optimal, but very specific, step-up procedure of Dunnett and Tamhane (1992).

1. Introduction.

Consider the multiple test problem where there are n hypotheses H_1, H_2, \dots, H_n , and corresponding P -values R_1, \dots, R_n , assuming the test statistics X_1, \dots, X_n are from a continuous distribution. Suppose that in a multiple test procedure the property

$$P(H_s, s \in I, \text{ are accepted} \mid H_s, s \in I, \text{ true}) \geq 1 - \alpha \quad (1)$$

Key words and phrases. multiple hypotheses, step-down tests, familywise error rate, Holm's procedure, Hochberg's procedure, multiple comparisons, P-value, correlation, Simes' test, Dunnett's test, multivariate t-distribution.

holds for prespecified size of test (familywise error rate: FWE) α , where I is any non-null subset of $\{1, 2, \dots, n\}$, and thus contains m items, $1 \leq m \leq n$. Then the procedure is said to control strongly the FWE. Let $R_{(1)}, R_{(2)}, \dots, R_{(n)}$ be the ordered P -values in ascending order, and $H_{(1)}, H_{(2)}, \dots, H_{(n)}$ the corresponding hypotheses.

To set this procedure in context : let $\Delta(p), 1 \leq p \leq n$ be a strictly increasing sequence of constants with $0 < \Delta(p) < 1, 1 \leq p \leq n$. A *step-down* procedure begins by testing if $R_{(1)} \leq \Delta(1)$. If so, reject $H_{(1)}$ and go on to test if $R_{(2)} \leq \Delta(2)$. If not, accept all hypotheses. In general, if $H_{(1)}, H_{(2)}, \dots, H_{(i)} \leq \Delta(i), 1 \leq i \leq j - 1$, then at step j the remaining hypotheses are $H_{(j)}, \dots, H_{(n)}$ and the inequality next to be checked is $R_{(j)} \leq \Delta(j)$. If it holds, reject $H_{(j)}$ and continue; otherwise accept $H_{(j)}, H_{(j+1)}, \dots, H_{(n)}$. The process may run at most until a decision is made on the basis of whether $R_{(n)} \leq \alpha$ or not.

A *step-up* testing algorithm begins by testing if $R_{(n)} \leq \Delta(n)$. If so, reject all hypotheses. If not, accept $H_{(n)}$, and go on to test if $R_{(n-1)} \leq \Delta(n - 1)$. In general, if $H_{(n-i+1)}, 1 \leq i \leq j$ have all been accepted, go on to test if $R_{(n-j)} \leq \Delta(n - j)$. If so, reject $H_{(n-j)}, H_{(n-j-1)}, \dots, H_{(1)}$; if not accept $H_{(n-j)}$, and continue.

The step-down procedure of Holm (1979) and the step-up procedure of Hochberg (1988) use the same set of cut-off constants:

$$\bar{\Delta}(p) = \alpha / (n - p + 1), \quad 1 \leq p \leq n. \quad (2)$$

Holm (1979) proved that with these constants the FWE is strongly controlled in the step-down procedure in general. That the same is true for the step-up (Hochberg) procedure for a wide range of joint distributions of X_1, X_2, \dots, X_n is a consequence of more recent papers of Sarkar and Chang (1997) and Sarkar (1999).

These papers are heavily focussed on an earlier procedure of Simes (1986) which influenced the considerations regarding the FWE in Hochberg (1988). Simes' (1986) procedure rejects any hypothesis $H_{(i)}$ corresponding to an $R_{(i)}$ for which $R_{(i)} \leq \alpha i / n, i = 1, \dots, n$, and is known to control the error rate (see Section 4) for a wide range of joint distributions. Although not really a step-down or step-up procedure,

it is closely related to the Holm and Hochberg procedures, and consequently it also figures in the sequel.

Generally, if we consider the same (but unspecific) $\Delta(i)$'s for a step-down and step-up procedure, the *probability of rejecting at least one hypothesis*, which shall be used as a measure of power in the sequel (see Section 4), for the step-down procedure is

$$P(R_{(1)} \leq \Delta(1)) , \quad (3)$$

whereas in the step-up procedure it is

$$1 - P(\text{Accept all } H_i) = 1 - P(R_{(n)} > \Delta(n), R_{(n-1)} > \Delta(n-1), \dots, R_{(1)} > \Delta(1)) , \quad (4)$$

and clearly (4) \geq (3). Moreover, since the step-up procedure rejects any hypothesis rejected by the step-down procedure, it is more powerful irrespective of the measure of power. However, one minor shortcoming even with the specific constants (2) is that there is no universal assurance that the FWE is controlled with the step-up (Hochberg) procedure.

Although it is less appealing as a measure of power, the *probability of rejecting precisely one hypothesis* for a step-down procedure is

$$P(R_{(1)} \leq \Delta(1), R_{(2)} > \Delta(2)) \quad (5)$$

while for a step-up procedure it is

$$P(R_{(n)} > \Delta(n), R_{(n-1)} > \Delta(n-1), \dots, R_{(2)} > \Delta(2), R_{(1)} \leq \Delta(1)) \quad (6)$$

so (5) \geq (6), which supports continuing interest in step-down procedures.

Although there has been a vast literature on the theme of stepwise hypothesis testing since Holm's (1979) paper (we cite only a selection within our References), the step-down and step-up procedures with constants (2) continue to be widely used because of their simplicity of application while satisfying the FWE property.

In a fundamental paper which heavily influenced the directions of our sequel, Dunnett and Tamhane (1992) reported, in comparing the Hochberg and Holm procedures under normal theory "... that step-up testing has non-negligible power advantage

only in those situations where most hypotheses are false and it is desired to reject all of them.” Their own procedure in this paper is step-up, and under normal theory in effect produces cut-off constants which satisfy the FWE requirement optimally, since the inequality (1) is satisfied with strict equality. Their constants require considerable computation, but are tabulated in a limited number of settings. These computational results provide a standard against which to compare.

In the present paper we are concerned with the practical aspects and effectiveness of a *step-down* procedure (labelled CS in the sequel) introduced in Seneta and Chen (1997, Section 6), which satisfies the FWE condition (1), *in general*, like Holm’s procedure. Write for the moment $R_{(i)} = R_{t_i}$, so t_i is random variable taking values from the set $\{1, 2, \dots, n\}$. Using the ordered P -values $R_{t_i}, i = 1, 2, \dots, n$ observed, define the index sets $K(\cdot)$ by $K(p) = \{t_p, t_{p+1}, \dots, t_n\}, p = 1, 2, \dots, n$ (these are random sets for $p \geq 2$) and write

$$\gamma(p) = \max_{(j \in K(p))} \sum_{i \in K(p)-j} P(R_i \leq \bar{\Delta}(p), R_j \leq \bar{\Delta}(p) | H_s, s \in K(p), \text{ true}) \quad (7)$$

for $1 \leq p \leq n - 1$, with $\gamma(n) = 0$. The $\bar{\Delta}(p)$ are given by (2). The cut-off constants for the CS procedure are then defined by

$$\tilde{\Delta}(p) = \frac{\alpha + \gamma(p)}{n - p + 1}. \quad (8)$$

They are therefore random for $2 \leq p \leq n - 1$ in general, and are of generally applicable form.

The CS step-down procedure is evidently more powerful than Holm’s since $\tilde{\Delta}(p) > \bar{\Delta}(p), 1 \leq p \leq n - 1$.

If X_i and X_j are independently distributed then

$$P(R_i \leq \bar{\Delta}(p), R_j \leq \bar{\Delta}(p)) = \bar{\Delta}^2(p) = \left(\frac{\alpha}{n - p + 1} \right)^2,$$

which for small α (say $\alpha = 0.05$) will be minuscule, and the contribution of terms of this kind to the sum in (7) will be negligible, and may be omitted, thus simplifying computation. This is a situation encountered in multiple comparisons (see Section 3).

On the other hand if X_i and X_j are perfectly positively correlated, then

$$P(R_i \leq \bar{\Delta}(p), R_j \leq \bar{\Delta}(p)) = P(R_i \leq \bar{\Delta}(p)) = \frac{\alpha}{n-p+1}$$

so the contribution to $\gamma(p)$ will be substantial. In fact, since then

$$\tilde{\Delta}(p) = \frac{\alpha + \gamma(p)}{n-p+1} = \alpha \left(\frac{2(n-p)+1}{(n-p+1)^2} \right)$$

$p = 1, 2, \dots, n$, and it can easily be checked that

$$\frac{1}{n-p+1} < \frac{1}{n-p} < \frac{2(n-p)+1}{(n-p+1)^2}$$

for $1 \leq p \leq n-2$, it follows that for tight positive correlation, the constants $\tilde{\Delta}(p)$ will be considerably higher than the corresponding Holm constants (2). This is illustrated in Section 4, Table 3.

Thus the constants $\tilde{\Delta}(p)$ of the CS procedure reflect the degree of positive association between the test statistics, the Holm constants (2) being the conservative boundary case obtained by taking all $\gamma(p) = 0$.

We shall include numerical evidence in Section 4 that the CS procedure is more powerful than the Hochberg step-up procedure which uses the simple constants $\bar{\Delta}(p)$ of (2).

In practice the extra computation involved in calculating the joint probabilities in (7) can be considerably lessened by applying the Holm step-down procedure (with the simple constants (2)) until for some p , $1 \leq p \leq n-1$, $R_{(p)} > \bar{\Delta}(p)$, or is very close to $\bar{\Delta}(p)$. Then $\gamma(p)$ may be computed for that specific p , to verify if the process is indeed to stop (if $R_{(p)} > \tilde{\Delta}(p)$) or to continue (if $R_{(p)} \leq \tilde{\Delta}(p)$) with $H_{(p)}$ rejected. If the p in question is $(n-1)$, only one calculation of a generalized cut-off $\tilde{\Delta}(n-1)$ is necessary, since $\bar{\Delta}(n) = \tilde{\Delta}(n) = \alpha$.

We thus envisage the CS procedure as an adjustment to Holm's step-down procedure, requiring modest additional calculation, and we illustrate by application in Section 3.

We now pass onto computational aspects.

2. Calculating Bivariate Probabilities.

Suppose Y_1, Y_2 are given by $Y_i = Z_i/\sqrt{\frac{C}{\nu}}, i = 1, 2$ where (Z_1, Z_2) is bivariate normal with $EZ_i = 0$, $\text{Var } Z_i = 1, i = 1, 2$; and $\text{Corr } (Z_1, Z_2) = \rho$, with (Z_1, Z_2) distributed independently of $C \sim \chi_\nu^2$. Thus (Y_1, Y_2) has joint bivariate t -distribution with parameters ρ and ν , and thus has pdf

$$f_{Y_1, Y_2}(y_1, y_2, \rho, \nu) = \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{y_1^2 - 2\rho y_1 y_2 + y_2^2}{\nu(1-\rho^2)}\right)^{-\frac{\nu}{2}-1}, \quad -\infty < y_1, y_2 < \infty. \quad (9)$$

Thus

$$\begin{aligned} & P(|Y_1| > x, |Y_2| > x) \\ &= P(Y_1 > x, Y_2 > x) + P(Y_1 > x, Y_2 < -x) \\ &\quad + P(Y_1 < -x, Y_2 < -x) + P(Y_1 < -x, Y_2 > x) \\ &= P(Y_1 > x, Y_2 > x) + P(Y_1 > x, -Y_2 > x) \\ &\quad + P(-Y_1 > x, -Y_2 > x) + P(-Y_1 > x, Y_2 > x). \end{aligned}$$

Putting $V_i = -Y_i, i = 1, 2$, since (V_1, V_2) has the same joint distribution as (Y_1, Y_2) , and since (Y_1, V_2) and (V_1, Y_2) have the same joint distribution (namely, the joint distribution of (Y_1, Y_2) but with the sign of the correlation coefficient reversed), the above is

$$P(|Y_1| > x, |Y_2| > x) = 2P(Y_1 > x, Y_2 > x) + 2P(Y_1 > x, V_2 > x). \quad (10)$$

Thus, in the hypothesis testing setup if

$$X_i = |Y_1| \quad \text{and} \quad X_j = |Y_2|$$

when $H_s, s \in I$, are true, then

$$P(R_i < \frac{\alpha}{n-p+1}, R_j < \frac{\alpha}{n-p+1} | H_s, s \in I) = P(|Y_1| > x, |Y_2| > x), \quad (11)$$

where

$$x = \Phi_\nu^{-1}\left(1 - \frac{\alpha}{2(n-p+1)}\right), \quad (12)$$

where Φ_ν is the cumulative distribution function of the t -distribution with ν degrees of freedom.

Thus given n, α, ν , and p , x is given by (12); and then with ρ specified by the experimental context for the given i and j , (11) is calculated from (10) and two numerical integrations

$$\begin{aligned} P(Y_1 > x, Y_2 > x) &= \int_x^A \int_x^A f(y_1, y_2; \rho, \nu) dy_1 dy_2, \\ P(Y_1 > x, V_2 > x) &= \int_x^A \int_x^A f(y_1, y_2; -\rho, \nu) dy_1 dy_2 \end{aligned} \quad (13)$$

where A is some large number. Such double integrations can be done on widely available packages such as MAPLE[©] 7 and MATHCAD[©], both of which have been used for the results reported in the sequel for such calculations.

3. Multiple Comparisons. An Application.

Suppose $X_{ij}, j = 1, \dots, J_i$ are independently and identically distributed $N(\mu_i, \sigma^2)$, and for $i = 1, 2, \dots, I$ form independent samples. Then $\bar{X}_i \sim N(\mu_i, \sigma^2/J_i)$, $i = 1, \dots, I$, and $SSE \sim \sigma^2 \chi_\nu^2$ are all independently distributed, where $N = \sum_{i=1}^I J_i$, $\nu = N - I$, and

$$SSE = \sum_{i=1}^I \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_i)^2,$$

using standard notation. Thus for $i \neq k, i, k = 1, 2, \dots, I$

$$\frac{\bar{X}_i - \bar{X}_k - (\mu_i - \mu_k)}{\sigma \sqrt{\frac{1}{J_i} + \frac{1}{J_k}}} \sim N(0, 1),$$

and so

$$T_{ik} = \frac{\bar{X}_i - \bar{X}_k - (\mu_i - \mu_k)}{S \sqrt{\frac{1}{J_i} + \frac{1}{J_k}}} \sim t_\nu$$

where $S^2 = SSE/\nu = MSE$. Also, Z_{ik} and $Z_{i\ell}, k \neq i, \ell \neq i, k \neq \ell$ are bivariate normal distributed with correlation coefficient

$$\rho = E(Z_{ik}, Z_{i\ell}) = \left(\left(1 + \frac{J_i}{J_k}\right) \left(1 + \frac{J_i}{J_\ell}\right) \right)^{-1/2}. \quad (14)$$

It follows that T_{ik} and $T_{i\ell}, k \neq i, \ell \neq i, k \neq \ell$, have a bivariate t -distribution with parameters $\nu = N - I$ and ρ given by (14).

Hence under the $n = \binom{I}{2}$ simultaneous hypotheses $H^{(i,k)} : \mu_i - \mu_k = 0$ for $1 \leq i < k \leq I$ the test statistics

$$\frac{\bar{X}_i - \bar{X}_k}{S\sqrt{\frac{1}{J_i} + \frac{1}{J_k}}}$$

have jointly a multivariate t -distribution with correlation parameters specified by (14), or by 0.

If $J_2 = J_k = J_\ell (= J, \text{ say})$, (14) becomes

$$\rho = \frac{1}{2}. \quad (15)$$

The Tukey T -procedure (studentized range test) for multiple comparisons which is appropriate in the case that all J_i are equal, $J_i = J$, $i = 1, \dots, I$, is encompassed by the statement that for fixed α , $0 < \alpha < 1$

$$P \left\{ \left| \frac{\bar{X}_i - \bar{X}_k - (\mu_i - \mu_k)}{S\sqrt{\frac{1}{J}}} \right| \leq Q_{\alpha, I, \nu}, 1 \leq i, k \leq I, i \neq k \right\} = 1 - \alpha, \quad (16)$$

where $Q_{\alpha, I, \nu}$ is the upper α -point of the studentized range distribution with parameter I , and $\nu = N - I = I(J - 1)$ degrees of freedom (Miller, 1981; 1997, pp.71-76).

When the J_i are unequal, the Tukey-Kramer procedure for multiple comparison is appropriate. Hayter (1984) has shown that

$$P \left\{ \left| \frac{\bar{X}_i - \bar{X}_k - (\mu_i - \mu_k)}{S\sqrt{\frac{1}{2} \left(\frac{1}{J_i} + \frac{1}{J_k} \right)}} \right| \leq Q_{\alpha, I, \nu}, 1 \leq i, k \leq I, i \neq k \right\} \geq 1 - \alpha,$$

where $\nu = N - I = \sum_{i=1}^I J_i - I$. This shows that the procedure for testing the $n = \binom{I}{2}$ simultaneous hypotheses for equality of pairwise means for a given overall nominal significance level α , is conservative in general, in the same sense as required by (1). On account of the equality with $(1 - \alpha)$ in (16), on the other hand, the Tukey

T -procedure may be thought in some sense most powerful in the case of equal sample sizes.

The actual methodology of application of the Tukey and Tukey-Kramer procedures for such multiple comparisons has been widely disseminated, for example by the popular textbook of Devore (2000, Section 10.2, pp. 414-418) whose notation we have (largely) adopted, and which contains a table (Table A.10) of the $Q_{\alpha, I, \nu}$ for analysis by hand-calculation. The procedures are available as widely used statistical packages whose use Devore (2000) also illustrates.

We compare the Tukey T -analysis in just one of the examples in Devore (2000, Example 10.5, pp.416-417) where all sample sizes are equal, with the outcomes of the Holm, Hochberg and CS procedures, with the CS procedure used as an aid to the Holm procedure. The origin of the example itself is a paper of Hattan and Eacho (1978) and concerns the effect of ethanol on sleep time in a sample of 20 rats ($N = 20$). There were 4 treatments (concentrations of ethanol; $I = 4$), each applied to 5 rats ($J = 5$). Thus $\nu = 16$. The values of the means n ascending numerical order are

$\bar{x}_4.$	$\bar{x}_3.$	$\bar{x}_2.$	$\bar{x}_1.$
32.760	47.920	61.540	79.280

and their differences are given in Table 1.

Table 1. Differences of Means

	$\bar{x}_2.$	$\bar{x}_3.$	$\bar{x}_4.$
$\bar{x}_1.$	17.74	31.36	46.52
$\bar{x}_2.$		13.62	28.78
$\bar{x}_3.$			15.16

Treatment 1 is the control (distilled water) and treatments 2, 3, 4 are increasing concentrations of ethanol. The $n = \binom{4}{2} = 6$ hypotheses being tested are $H^{(i,j)} : \mu_i - \mu_j = 0, i < j$. Testing at $\alpha = 0.05$ level, $Q_{0.05, 4, 16} = 4.046$. The observed

value of S^2 is $s^2 = 92.96250$ so $s = 9.641706$; and $Q_{0.05,4,16} \sqrt{s^2/J} = 17.45$. Devore (2000, p.417) writes, consequently: “The interpretation ... must be done with care, since we seem to have concluded that treatments 2 and 3 do not differ, 3 and 4 do not differ, yet 2 and 4 do differ ... Treatment 1 has significantly higher true average sleep REM time than any of the treatments.” Hattan and Eacho’s (1978, p.842) conclusions, however, although details are not supplied, are that treatment 4 is significantly more potent than treatment 3, which is significantly more potent than treatment 2, with all 3 dosages differing significantly from the control.

The t_{16} -scores are given by Table 2, and the corresponding P -values by Table 3.

Table 2.

t_{16} scores. $(\bar{x}_i - \bar{x}_j)/(s\sqrt{2/J})$. Rats.

$i \setminus j$	2	3	4
1	2.909	5.143	7.629
2		2.234	4.720
3			2.486

Table 3.

P -values. $2P(t_{16} > \text{Table 2})$. Rats.

$i \setminus j$	2	3	4
1	0.01025	9.819×10^{-5}	0.1021×10^{-5}
2		0.04011	23.12×10^{-5}
3			0.02435

Ranking hypotheses in increasing order of P -values we have

$$H^{(1,4)}, H^{(1,3)}, H^{(2,4)}, H^{(1,2)}, H^{(3,4)}, H^{(2,3)}. \quad (17)$$

The values of $\bar{\Delta}(p) = \alpha/(n + p - 1) = 0.05/(7 - p)$, for $p = 1, 2, 3, 4, 5, 6$ are

$$8.333 \times 10^{-3}, 0.0100, 0.01250, 0.01667, 0.0250, 0.05 .$$

Thus with the step-down procedure using Holm's constants (2), all hypotheses are rejected. The conclusion that *all* treatments differ significantly makes rather more sense intuitively for this experimented setting than that obtained with Tukey's test, in which the single value $Q_{0.05,4,16}$ is used as cut-off for standardized values and only $H^{(1,4)}, H^{(1,3)}, H^{(2,4)}$ and $H^{(1,2)}$ are rejected. This is somewhat akin to using the same "Bonferroni adjustment" cut-off value $\alpha/n = 0.05/6 = 8.333 \times 10^{-3}$ for each P -value. That would result in the rejection only of $H^{(1,4)}, H^{(1,3)}, H^{(2,4)}$.

Some reassurance is useful over the rejection of $H^{(3,4)}$ in the simple step-down procedure. The CS-procedure increases the cut-off $\bar{\Delta}(5) = 0.05/2 = 0.025$ for the realized P -value of 0.02435 to a cut-off of $\tilde{\Delta}(5) = 0.0274$ for this random experiment. We set out the details of this calculation now.

Let $H_1 = H^{(1,2)}, H_2 = H^{(1,3)}, H_3 = H^{(1,4)}, H_4 = H^{(2,3)}, H_5 = H^{(2,4)}, H_6 = H^{(3,4)}$. Then in the notation of our Section 1, $K(5) = \{t_5, t_6\} = \{6, 4\}$, from the ranking (17). Thus (7) becomes

$$\gamma(5) = P(R_6 \leq \bar{\Delta}(5), R_4 \leq \bar{\Delta}(5)).$$

Since $\bar{X}_3, -\bar{X}_4$ and $\bar{X}_2, -\bar{X}_3$ are correlated with correlation $\rho = -1/2$ we need to calculate x in (12) $\nu = 16$, and then (11) from (10) and (13), using (9) with $\nu = 16$ and $\rho = 1/2$. Thus

$$x = \Phi_{16}^{-1}\left(1 - \frac{0.05}{4}\right) = \Phi_{16}^{-1}(0.9875) = 2.472878 \dots$$

$$\gamma(5) = 2(0.002268828200 + 0.00009509894086) = 0.004729634$$

$$\tilde{\Delta}(5) = \frac{0.05 + 0.0047296}{2} = 0.0274 .$$

Since all hypotheses are rejected by Holm's step-down procedure, using the same cut-offs $\bar{\Delta}(p), p = 1, 2, \dots, n$ in Hochberg's step-up procedure gives the same conclusion. The example also questions the power of Tukey's T-test as compared to simple step-wise procedures; this may be addressed more generally elsewhere.

4. Tabulation and Simulation Investigations.

We consider upper-tail tests where $X_i = |T_i|, i = 1, 2, \dots, n$, with T_1, T_2, \dots, T_n defined by $T_i = W_i/\sqrt{C/\nu}, i = 1, \dots, n$ where the W_i 's are multivariate normal with $EW_i = \nu_i, \text{Var}(W_i) = 1, i = 1, \dots, n, \text{Corr}(W_i, W_j) = \rho, i \neq j,$ and are independently distributed of $C \sim \chi_\nu^2$.

Thus under $H_i : \mu_i = 0, i = 1, \dots, n$, the $T_i, i = 1, \dots, n$ have jointly a multivariate exchangeable t-distribution with parameters $n, \rho, (\rho \geq -1/(n-1)), \nu$, as in Dunnett's tests for comparison of the means of n treatment normal samples with a control sample, all samples being independent. This is a simpler setting than the one considered in Section 3 but of continuing interest and publishing activity. In particular the cut-offs $\tilde{\Delta}(p), p = 1, \dots, n$ are *non-random* in this setting.

It will be useful to a reader applying the CS procedure in this setting to have available the constants $\tilde{\Delta}(p)$ for various n and p , and we provide a tabulation in the limiting case $\nu = \infty$ (that is, with $T_i = W_i, i = 1, 2, \dots, n$ multivariate normal). Writing $m = n - p + 1$, our tabulation (Table 4) is of $\tilde{\Delta}'(m), m = 1, 2, \dots, 8$, where

$$\tilde{\Delta}'(m) = \tilde{\Delta}(n - p + 1)$$

for various ρ . The first line gives α/m , the Holm constants.

Table 4

$$\tilde{\Delta}'(m), \nu = \infty, \alpha = 0.05.$$

m	1	2	3	4	5	6	7	8
α/m	0.05	0.025	0.16667	0.01250	0.01	0.00833	0.00714	0.00625
0.0	0.05	0.02531	0.01685	0.01262	0.01008	0.00839	0.00718	0.00628
0.1	0.05	0.02537	0.01689	0.01265	0.01010	0.00841	0.00720	0.00630
0.3	0.05	0.02589	0.01723	0.01290	0.01030	0.00857	0.00733	0.00641
0.5	0.05	0.02676	0.01801	0.01351	0.0108	0.00898	0.00768	0.00670
0.6	0.05	0.02751	0.01864	0.01403	0.01122	0.00933	0.00799	0.00697
0.7	0.05	0.02851	0.01951	0.01475	0.01182	0.00985	0.00844	0.00737
0.8	0.05	0.02989	0.02073	0.01578	0.01270	0.01061	0.00910	0.00796
0.9	0.05	0.03192	0.02259	0.01736	0.01406	0.01180	0.01016	0.00891
0.95	0.05	0.03349	0.02403	0.01861	0.01515	0.01276	0.01101	0.00968
0.99	0.05	0.03568	0.02608	0.02039	0.01670	0.01413	0.01224	0.01079
1.00	0.05	0.03750	0.02778	0.02187	0.018	0.01528	0.01327	0.01172

Error Rate

This is defined as

$$P(\text{Reject at least one } H_i, i = 1, \dots, n | \mu_i = 0, i = 1, \dots, n)$$

and is most easily obtained for given n, ν and $\rho < 1$ by simulation. Our investigation was for $n = 3$ and $\alpha = 0.05$. 20,000 correlated triples, each an independent observation on (T_1, T_2, T_3) where $\mu_i = 0, i = 1, 2, 3$, were produced at each ρ from a set of 4 quadruples which are observations on (S_1, S_2, S_3, S_4) where the S'_i 's were i.i.d. $N(0, 1)$, with

$$W_i = \rho^{1/2} S_1 + (1 - \rho)^{1/2} S_{i+1}, \quad i = 1, 2, 3.$$

When $\nu = \infty$, $W_i = T_i, i = 1, 2, 3$. Thus the random number generator produced in effect 4 columns of 20,000 observations each on independent $N(0, 1)$ random

variables. The statistical package used was SPIDA[©] Version 6.21 (1 January, 1999); (SPIDA=Statistical Package for Interactive Data Analysis, Statistical Computing Laboratory, Macquarie University, Sydney, Australia).

For ν finite, we considered only $\nu = 16$. In this situation, with every S_1, S_2, S_3, S_4 generated, further readings on S_5, S_6, \dots, S_{20} were generated, and C evaluated from

$$C = S_5^2 + \dots + S_{20}^2 .$$

The error rate for each of the procedures Holm, Hochberg, and CS was computed from the same set of 20,000 independent triples of observations on (T_1, T_2, T_3) for each value of ρ . The values shown in Tables 5 and 6, calculated for each procedure from the same 20,000 triples, at least give the *relativities* of the 3 procedures, although the estimates of the probability p_e , the error rate, are likely accurate only to the second decimal place, since the standard deviation in estimating p_e by a relative frequency is at most $\sqrt{1/(4 \times 20,000)} = 3.535 \times 10^{-3}$.

At $\rho = 1$ the error rate was obtained by calculation.

The error rate for the Dunnett and Tamhane (1992) procedure as already mentioned is exactly the nominal $\alpha = 0.05$.

Table 5

Estimated Error Rate. $n = 3, \nu = 16, \alpha = 0.05$.

ρ	0.0	0.1	0.3	0.5	0.6	0.7	0.8	0.9	0.95	0.99	1.00
Holm	0.0486	0.0469	0.0448	0.0430	0.0412	0.0380	0.0340	0.0291	0.0245	0.0193	0.0167
Hochberg	0.0489	0.0472	0.0455	0.0442	0.0428	0.0397	0.0371	0.0347	0.0346	0.0411	0.0500
CS	0.0506	0.0485	0.0475	0.04735	0.0466	0.0449	0.0423	0.0388	0.0365	0.0321	0.0278
Simes	0.0498	0.0480	0.0464	0.0455	0.0446	0.0415	0.0393	0.0375	0.0366	0.0417	0.0500

Table 6Estimated Error Rate. $n = 3$, $\nu = \infty$, $\alpha = 0.05$.

ρ	0.0	0.1	0.3	0.5	0.6	0.7	0.8	0.9	0.95	0.99	1.00
Holm	0.0487	0.0498	0.0485	0.0457	0.0441	0.0400	0.0365	0.0308	0.0268	0.0214	0.0167
Hochberg	0.0490	0.0498	0.0489	0.0464	0.0454	0.0422	0.0393	0.0355	0.0347	0.0409	0.0500
CS	0.0493	0.0504	0.0502	0.0491	0.0480	0.0458	0.0444	0.0397	0.0366	0.0314	0.0278
Simes	0.0496	0.0504	0.0496	0.0483	0.0469	0.0442	0.0412	0.0379	0.0373	0.0414	0.0500

Notice that the error rate for the Holm procedure:

$$P(R_{(1)} \leq \bar{\Delta}(1))$$

(see (3)) is the same as for the older ‘‘Bonferroni adjustment’’ procedure which uses $\Delta(p) = \alpha/n$ for each $p = 1, 2, \dots, n$.

Tables 5 and 6 show that the CS procedure outperforms, as is to be expected inasmuch as it takes into account association between test statistics, the Hochberg and Simes procedures up to quite high values of ρ . The reader will notice the occasional simulated error rate exceeding 0.05 due to sampling fluctuation; and the increase in the error rate up to $\alpha = 0.05$ for ρ very close to 1.

To investigate power beyond the error rate as measure of power, we report only on the normal case ($\nu = \infty$), with $\rho = 0.5, 0.7, 0.9$, and take as our measure of power

(A) Probability of rejecting at least one $H_i : \mu_i = 0, i = 1, 2, \dots, n$ when in fact $\mu_i = \delta, i = 1, 2, \dots, n$.

(B) Probability of rejecting precisely one of $H_i : \mu_i = 0, i = 1, 2, \dots, n$, when in fact $\mu_1 = \delta, \mu_2 = 0, \dots, \mu_n = 0$.

The choices of δ below are influenced by the numerical investigations in Simes (1986). The simulated observation triples are now on

$$T_i = W_i = \rho^{1/2} S_1 + (1 - \rho)^{1/2} S_{i+1} + \mu_i^0, \quad i = 1, 2, \dots, n$$

where μ_i^0 is the value of μ_i specified in place of that given by H_i . The values in the DT column in Tables 7 - 9 come from using the (optimal) cut-off constants of

Dunnett and Tamhane (1992), Table 2. These are not available in that paper for ρ greater than 0.5. Several common applications use the value $\rho = 0.5$ (e.g. Dunnett's tests with equal sample sizes).

For $n = 3$ the simulations produced 20,000 observations of the quadruples (S_1, S_2, S_3, S_4) .

Table 7.

Estimated Probability of Rejection. $n = 3, \nu = \infty, \alpha = 0.05, \delta = 2$.

ρ	(A)			(B)		
	Hochberg	CS	DT	Hochberg	CS	DT
0.5	0.6028	0.6042	0.6174	0.3426	0.3542	0.3577
0.7	0.5478	0.5567	—	0.3395	0.3605	—
0.9	0.4819	0.4993	—	0.3391	0.3836	—

Table 8

Estimated Probability of Rejection. $n = 3, \nu = \infty, \alpha = 0.05, \delta = 3$.

ρ	(A)			(B)		
	Hochberg	CS	DT	Hochberg	CS	DT
0.5	0.9241	0.9227	0.9297	0.7129	0.7204	0.7234
0.7	0.8878	0.8909	—	0.7168	0.7354	—
0.9	0.8385	0.8596	—	0.7196	0.7725	—

Table 9

Estimated Probability of Rejection. $n = 3, \nu = \infty, \alpha = 0.05, \delta = 4$.

ρ	(A)			(B)		
	Hochberg	CS	DT	Hochberg	CS	DT
0.5	0.9959	0.9956	0.9963	0.9123	0.9126	0.9133
0.7	0.9900	0.9900	—	0.9202	0.9226	—
0.9	0.9789	0.9795	—	0.9296	0.9380	—

Finally we report some simulation results akin to Tables 7 to 9 in the case $n = 5$ (still with $\nu = \infty, \alpha = 0.05$) to display the effect of larger n . Here we began with 10,000 observations on the sextuples

$$(S_1, S_2, \dots, S_6), \text{ with } T_i = W_i = \rho^{1/2} S_1 + (1 - \rho)^{1/2} S_{i+1} + \mu_i^0, i = 1, 2, \dots, 5, .$$

With $n = 10,000$ independent replications, the standard error of estimate of the probability p_e is at most $1/\sqrt{4 \times 10,000} = 5 \times 10^{-3}$, so the estimates are likely correct to 2 decimal places, but we have given more to show the relativities between the estimates produced from the same 10,000 observations on the quintuples $(T_1, T_2, T_3, T_4, T_5)$.

The statistical package used to produce Tables 7-12 including random number generation, and use of macros (especially in the case $n = 5$) was MINITAB[®] Release 9.2.

Table 10

Estimated Probability of Rejection. $n = 5, \nu = \infty, \alpha = 0.05, \delta = 2$.

ρ	(A)			(B)		
	Hochberg	CS	DT	Hochberg	CS	DT
0.5	0.6126	0.6206	0.6416	0.2789	0.2867	0.2980
0.7	0.5337	0.5522	—	0.2702	0.2906	—
0.9	0.4401	0.4629	—	0.2670	0.3077	—

Table 11

Estimated Probability of Rejection. $n = 5, \nu = \infty, \alpha = 0.05, \delta = 3$.

ρ	(A)			(B)		
	Hochberg	CS	DT	Hochberg	CS	DT
0.5	0.9349	0.9360	0.9411	0.6532	0.6614	0.6726
0.7	0.8873	0.8951	—	0.6547	0.6759	—
0.9	0.8133	0.8294	—	0.6527	0.6976	—

Table 12

Estimated Probability of Rejection. $n = 5, \nu = \infty, \alpha = 0.05, \delta = 4$.

ρ	(A)			(B)		
	Hochberg	CS	DT	Hochberg	CS	DT
0.5	0.9965	0.9967	0.9973	0.8934	0.8941	0.8954
0.7	0.9875	0.9883	—	0.9040	0.9063	—
0.9	0.9739	0.9757	—	0.9106	0.9251	—

5. Conclusion.

A reasonable inference to the above tabulations is that in normal theory settings, where the bivariate normal distribution has known correlation coefficient, and the number of hypotheses, n , is not large, a CS cut-off may be usefully invoked as an adjustment to the simple Holm procedure, for quick assessment. More generally, it may be useful to incorporate the CS cut-off into computer package procedures used for multiple comparisons, although the power relative to the Tukey-T and Tukey-Kramer procedures needs further investigation.

Acknowledgement.

Much of the computational work was done by ES during Fall Semesters within the period 1999-2002 at the Department of Mathematics, University of Virginia, whose hospitality is gratefully acknowledged. He also thanks Petra Macaskill, University of Sydney, for computations of bivariate probabilities, with MATHCAD[®] during this period.

References.

- Chen, J.T. and Seneta, E. (1999) A stepwise rejective test procedure with strong control of familywise error rate. *Bulletin of the International Statistical Institute*, ISI 99, 52nd Session. Contributed papers. Tome LVII, Three Books. Book 3. Helsinki, 1999. pp.241-242.
- Devore, J.L. (2000) *Probability and Statistics for Engineering and the Sciences*. Duxbury, Pacific Grove, CA.
- Dunnett, C.W. and Tamhane, A.C. (1992) A step-up multiple test procedure. *J. American Statistical Association*, **87**, 162-170.
- Hattan, D.G. and Eacho, P.I. (1978) Relationship of ethanol blood level to REM and non-REM sleep time and distribution in the rat. *Life Sciences*, **22**, 839-846.
- Hayter, A.J. (1984) A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics*, **12**, 61-75.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800-802.

- Hochberg, Y. and Rom, D. (1995). Extensions of multiple testing procedures based on Simes' test. *J. Statist. Plan. Inf.* **48**, 141-152.
- Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparison Procedures*. New York: Wiley.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.
- Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75**, 383-386.
- Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika* **76**, 624-625.
- Miller, R.G. (1997) *Beyond ANOVA, Basics of Applied Statistics*. Chapman and Hall, London.
- Miller, R.G. (1981) *Simultaneous Statistical Inference*. 2nd Edn. Springer, New York.
- Miller, R.G. (1985) Multiple comparisons. *Encyclopedia of Statistical Sciences*. (S. Kotz and N.L. Johnson, ed.) **5**. pp.679-689. , Wiley, New York.
- Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663-666.
- Samuel-Cahn, E. (1996). Is Simes improved Bonferroni procedure conservative? *Biometrika* **83**, 929-933.
- Sarkar, S.K., (1988). Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *Ann. Statist.*, **26**, 494-504.
- Sarkar, S.K. and Chang, C.K. (1997) The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.*, **92**, 1601-1608.
- Seneta, E. (1993). Probability Inequalities and Dunnett's Test. In F.M. Hoppe, Ed: *Multiple Comparisons, Selection, and Applications in Biometry*, Chapter 3, pp 29-45, New York: Marcel Dekker.

- Seneta, E. and Chen, T. (1997) A sequentially rejective test procedure. *Theory of Stochastic Processes*, **3 (19)**, No. 3-4, 393-402 (TBIMC Scientific Publishers, Kyiv, Ukraine.)
- Shaffer, J.P. (1986). Modified sequentially rejective multiple test procedure. *J. American Statistical Association* **81**, 826-831.
- Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751-754.
- Stoline, M.R. (1983). The Hunter method of simultaneous inference and its recommended use for applications having large known correlation structures. *Journal of the American Statistical Association* **78**, 366-370.
- Wright, S.P. (1992). Adjusted P-Values for simultaneous inference. *Biometrics* **48**, 1005-1013.