Biometrical Journal **43** (2001) 5, 657–663

Obtaining Critical Values for Simultaneous Confidence Intervals and Multiple Testing

PAUL N. SOMERVILLE

Department of Statistics
University of Central Florida
USA

FRANK BRETZ

Department of Bioinformatics
University of Hannover
Germany

Summary

There are many situations where it is desired to make simultaneous tests or give simultaneous confidence intervals for linear combinations (contrasts) of population or treatment means. SOMERVILLE (1997, 1999) developed algorithms for calculating the critical values for a large class of simultaneous tests and simultaneous confidence intervals. Fortran 90 and SAS-IML batch programs and interactive programs were developed. These programs calculate the critical values for 15 different simultaneous confidence interval procedures (and the corresponding simultaneous tests) and for arbitrary procedures where the user specifies a combination of one and two sided contrasts. The programs can also be used to obtain the constants for "step-down" testing of multiple hypotheses. This paper gives examples of the use of the algorithms and programs and illustrates their versatility and generality. The designs need not be balanced, multiple covariates may be present and there may be many missing values. The use of multiple regression and dummy variables to obtain the required variance covariance matrix is illustrated. Under weak normality assumptions the methods are "exact" and make the use of approximate methods or "simulation" unnecessary.

Key words: Multiple testing; Multiple comparisons; Simultaneous confidence intervals; Critical values.

1. Methodology

Assume we have k populations or treatments with unknown means $\mu_1, \mu_2, \ldots, \mu_k$. Let $\boldsymbol{x}^T = (x_1, x_2, \ldots, x_k)$ be an estimate of $\mu^T = (\mu_1, \mu_2, \ldots, \mu_k)$ having the multivariate normal distribution (MVN) with mean μ and variance-covariance matrix $\sigma^2\Sigma$ with Σ known and s^2 the usual estimate of σ^2 with μ degrees of freedom. Let $\boldsymbol{c} = (c_1, c_2, \ldots, c_k)$ where $c_1 + c_2 + \ldots + c_k = 0$. It is often desired to simultaneously test a set of m hypotheses or equivalently obtain a set of m simultaneous

 0323-3847/01/0509-0657 $ 17.50+.50/0

confidence intervals. Since the composite hypothesis H: $\boldsymbol{c}_i\mu \neq 0$ is equivalent to the individual hypotheses $-\boldsymbol{c}_i\mu < 0$ and H: $-\boldsymbol{c}_i\mu < 0$, we limit our discussion to one sided hypotheses of the type H: $\boldsymbol{c}_i\mu < 0$. Corresponding to each set of hypotheses, we define a set $\boldsymbol{B}$ which is composed of $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_m$. The programs can solve for the value of q such that

$$\Pr\left[\boldsymbol{c}_i\boldsymbol{x}/(\operatorname{var}(\boldsymbol{c}_i\boldsymbol{x}))^{1/2} < q/2^{1/2}\right] = 1 - \alpha \qquad \boldsymbol{c}_i \in \boldsymbol{B}$$

for arbitrary values for $\boldsymbol{c}_i$, m, k, ν, Σ (or the correlation matrix), and $\alpha > .5$. The quantity $\operatorname{var}(\boldsymbol{c}_i\boldsymbol{x})$ is the variance of the contrast $\boldsymbol{c}_i\boldsymbol{x}$, estimated, if necessary, using the appropriate degrees of freedom, and $q/2^{1/2}$ is the "critical value" for the statistical test. We use $q/2^{1/2}$ rather than q, consistent with common usage for the Tukey multiple comparison procedure. The critical values q produced by the program will thus be larger by a factor of $2^{1/2}$ than the published results for some of the other simultaneous confidence interval procedures. The program uses a combination of Monte Carlo and quadrature. The methodology is described in SOMERVILLE (1999). For 15 simultaneous confidence interval procedures the programs internally generate the appropriate sets $\boldsymbol{B}$ and the integer m, and calculate the value of q corresponding to specified values of α, ν and k.

For all of these procedures, arbitrary nonsingular values of Σ (or the correlation matrix) may be assigned. This implies that the $\boldsymbol{x}$ estimates may be obtained from designs with unequal sample sizes, missing values, unequal variances and that the estimates may be correlated. It should be emphasized that the correlation matrix contains the correlations between the individual estimates x_i and not between the contrast estimates $\boldsymbol{c}_i\boldsymbol{x}$. The procedures are: 1: all-pairs pairwise comparisons (TUKEY, 1953), 2: ordered pairwise comparisons (BOFINGER, 1985), 3: multiple comparisons with the best (HSU, 1984), 4: one-sided comparisons with a control (DUNNETT, 1955), 5: two-sided comparisons with a control (DUNNETT, 1964), 6: one-sided ordered comparisons (HAYTER, 1990), 7: one-sided successive ordered treatments (LIU et al., 1999), 8: two-sided successive ordered treatments (LIU et al., 1999), 9: umbrella contrasts, 10: generalized umbrella contrasts, 11: Williams type contrasts (BRETZ, 1999), 12: Marcus type contrasts (BRETZ, 1999), 13: step contrasts (HIROTSU, 1997), 14: order-constrained contrasts (MCDERMOTT and MUDHOLKER, 1993), and 15: isotonic contrasts (BRETZ, 1999).

The programs have two other capabilities. For an arbitrary set $\boldsymbol{B}$, specified by a user, and arbitrary values for k, ν, Σ and α, the programs calculate q (procedure 0.). Thus critical values for testing arbitrary sets of hypotheses of the type H: $\boldsymbol{c}_i\mu < 0$, can be obtained. In addition, suppose we replace m_1 of the hypotheses by the condition $\boldsymbol{c}_i\boldsymbol{x}(\operatorname{var}(\boldsymbol{c}_i\boldsymbol{x}))^{1/2} < q_1/2^{1/2}$. Again, for k, ν, Σ and α, the programs can calculate q (procedure 20). This is equivalent to

$$\Pr\left[\boldsymbol{c}_i\boldsymbol{x}/(\operatorname{var}(\boldsymbol{c}_i\boldsymbol{x}))^{1/2} < q/2^{1/2}, \qquad i = 1, 2 \ldots, m - m_1\right.$$

and

$$\left.\boldsymbol{c}_i\boldsymbol{x}/(\operatorname{var}(\boldsymbol{c}_i\boldsymbol{x}))^{1/2} < q_1/2^{1/2}, \qquad i = m - m_1 + 1, \ldots, m\right] = 1 - \alpha\,.$$

The latter procedure has been used by SOMERVILLE, LIU, MIWA, and HAYTER (2001). It may be noted that the set $\boldsymbol{B}$ for each of procedures 2 to 10 is a subset of that for the Tukey procedure. More details are given in SOMERVILLE (1997, 1999).

2. A Multifactorial Dose-Response Study Example

Table 1 gives the data (fictitious) from an experiment to compare the efficacy EF of 7 dosage levels L of a drug on 40 patients. The experiment was conducted using 3 different hospitals H, both sexes S. Data for two covariates D and E were recorded at the beginning of the experiment. The object of the analysis was to determine if one of the two dosage levels 4 or 5 could be the peak level.

Dummy (indicator) variables for the qualitative variables were coded as follows:

$x_i = 1$ if dose level i was administered, otherwise $x_i = 0$, for $i = 1, 2, \ldots, 7$

$h_2 = 1$ if hospital 2 was used, otherwise $h_2 = 0$

$h_3 = 1$ if hospital 3 was used, otherwise $h_3 = 0$

$s_2 = 1$ patient was F, otherwise $s_2 = 0$.

The regression model was

$$y = \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3 + \beta_{14}x_4 + \beta_{15}x_5 + \beta_{16}x_6 + \beta_{17}x_7 + \beta_{22}h_2 + \beta_{23}h_3 + \beta_{32}s_2 + \beta_4 D + \beta_5 E + \varepsilon$$

(note the absence of an intercept!). The estimates of the efficacy for the 7 dosage levels (adjusted for hospitals, sex, and covariates D and E, and estimated by b_{11}, b_{12}, …, b_{17}) were 34.48 (level 1), 40.54 (level 2), 41.06 (level 3), 46.49 (level 4), 50.01 (level 5), 42.02 (level 6), and 39.33 (level 7). The associated variance covar-

Table 1
Data from Experiment 1

EF	L	H	S	D	E	EF	L	H	S	D	E	EF	L	H	S	D	E	EF	L	H	S	D	E
34	1	1	1	10	04	49	4	2	2	21	45	38	3	2	1	40	41	51	5	3	2	20	80
37	1	2	2	15	50	44	4	2	1	30	40	38	3	2	1	10	32	42	6	1	1	10	50
30	1	3	1	20	40	49	4	2	2	25	65	43	3	2	2	18	60	45	6	2	2	30	90
37	1	2	2	21	60	46	5	3	1	26	60	43	3	2	2	16	70	38	6	3	1	35	70
34	1	1	1	21	70	51	5	3	2	15	75	41	3	3	2	18	80	47	6	1	2	26	40
42	2	1	1	30	80	46	5	3	1	12	45	41	3	3	2	20	85	42	6	1	1	27	60
42	2	1	1	20	35	51	5	3	2	17	55	46	4	1	1	14	35	42	7	2	2	35	65
40	2	2	2	10	40	50	5	1	1	15	65	51	4	1	2	20	50	37	7	2	1	28	70
40	3	1	1	18	70	53	5	2	2	30	70	46	4	1	1	31	41	40	7	3	2	30	90
40	3	1	1	20	65	46	5	3	1	40	65	46	4	1	1	40	45	35	7	3	1	31	45

iance matrix (upper left hand corner of the of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ matrix) was

0.53216	0.42662	0.53323	0.39077	0.45658	0.44176	0.52788
0.42662	0.93232	0.79182	0.56392	0.55381	0.59180	0.67782
0.53323	0.79182	1.49670	0.84864	0.63247	0.70566	0.88876
0.39077	0.56392	0.84864	0.70872	0.48012	0.52985	0.63036
0.45658	0.55381	0.63247	0.48012	0.81523	0.61291	0.73320
0.44176	0.59180	0.70566	0.52985	0.61291	0.81588	0.69811
0.52788	0.67782	0.88876	0.63036	0.73320	0.69811	1.10070

The estimate of σ^2 was 0.467037 with 28 df. Using QBATCH4.FOR or QBATCH4.SAS, for procedure 10. we obtain $q = 4.007$ for $\alpha = .05$ using the above variance covariance matrix. The confidence interval for $(\mu_4 - \mu_3)$ is $(4.440, \infty)$ calculated as follows:

$$(46.49 - 41.06) - [(0.70872 + 1.49670 - 2 \times 0.84864 \times 0.467037]^{1/2} \times 4.007 \cdot 2^{1/2} < \mu_4 - \mu_3 < \infty .$$

If prior to looking at the data we had wished to test that the peak dosage level was level 5, using procedure 9, we would have calculated a q value of 3.904, and the confidence intervals for $(\mu_5 - \mu_1)$, $(\mu_5 - \mu_2)$, $(\mu_5 - \mu_3)$, $(\mu_5 - \mu_4)$, $(\mu_5 - \mu_6)$ and $(\mu_5 - \mu_7)$. Table 2 contains the complete sets of simultaneous confidence intervals for both procedures 9 and 10. Procedure 10 allows us to conclude that levels 4 or 5 represent the peak, while procedure 9 allows us to say it is 5.

Table 2
Simultaneous confidence intervals for procedures 9 and 10

	Proc. 10	Proc. 9		Proc. 10	Proc. 9
$\mu_4 - \mu_1$	$(11.418, \infty)$		$\mu_5 - \mu_2$	$(8.231, \infty)$	$(8.263, \infty)$
$\mu_4 - \mu_2$	$(4.957, \infty)$		$\mu_5 - \mu_3$	$(6.916, \infty)$	$(6.968, \infty)$
$\mu_4 - \mu_3$	$(4.440, \infty)$		$\mu_5 - \mu_4$		$(2.456, \infty)$
$\mu_4 - \mu_6$	$(3.572, \infty)$		$\mu_5 - \mu_6$	$(7.206, \infty)$	$(7.227, \infty)$
$\mu_4 - \mu_7$	$(6.100, \infty)$		$\mu_5 - \mu_7$	$(9.811, \infty)$	$(9.811, \infty)$
$\mu_5 - \mu_1$	$(14.986, \infty)$	$(15.008, \infty)$			

3. Some Program Results

If treatment or mean estimates are correlated, the values of q (and the width of the associated confidence intervals) may differ significantly from the uncorrelated case. To illustrate, q-values were computed for procedures 1 to 10, for $k = 7$, $\alpha = .01$ and $\nu = 28$, and for both the uncorrelated case, and when the variance covariance of the estimates is given by Table 3.

Table 3
Variance Covariance matrix for example in Section 3

7.760	0.555	12.063	−0.277	− 1.106	− 2.489	− 4.425
0.555	8.244	−15.363	1.784	7.139	16.061	28.553
12.063	−15.363	58.322	−4.105	−16.424	−36.951	−65.693
−0.277	1.784	− 4.105	0.766	1.911	3.108	6.966
−1.106	7.139	−16.424	1.911	6.855	14.594	26.929
−2.489	16.061	−36.951	3.108	14.594	35.090	59.596
−4.425	28.553	−65.693	6.966	26.929	59.596	107.184

Table 4 gives selected q values The first value is for the "balanced case" (Σ is the identity matrix), while the value in parenthesis is the corresponding value when the estimates $\boldsymbol{x}$ of the population or treatment means μ are correlated. MCB-i refers to the q value used when making comparisons with treatment i for HSU's "Multiple Comparisons with the Best". DUN(1)-i and DUN(2)-i refers to the 1- and 2-sided Dunnett procedures when the population i is the control population. In a similar fashion UMB-i refers to the case where population i is the peak population and UMB/$i-j$ refers to the case where the peak population is assumed to be somewhere in the interval (i, j). It is worth noting that the q-values for the correlated case are as much as 15% below the corresponding values for the balanced case.

For the cases in Table 4, the Fortran 90 program QBATCH4 was used to obtain the actual p-values when the value of the largest calculated q value was equal to the critical value for the uncorrelated case. The p-values ranged from 0.0033 to 0.0091, as contrasted with the intended value of 0.01.

Table 4
Selected q-values for the example

1: TUKEY	5.441(4.805)	5: DUN(2)-2	4.826(4.626)	9: UMB/4−7	5.227(4.627)
2: BOF	5.102(4.557)	5: DUN(2)-3	4.826(4.186)	9: UMB/6−7	5.127(4.597)
3: MCB-2	4.424(4.371)	6: HAYT	5.102(4.557)	9: UMB/5−6	5.026(4.590)
3: MCB-3	4.424(3.764)	7: SOT(1)	4.826(4.546)	9: UMB/1−4	5.227(4.695)
3: MCB-5	4.424(4.360)	8: SOT(2)	4.876(4.482)	9: UMB/2−4	5.101(4.612)
4: DUN(1)-1	4.424(4.191)	9: UMB-3	4.868(4.260)	9: UMB/1−3	5.170(4.611)
4: DUN(1)-2	4.424(4.371)	9: UMB-4	4.831(4.403)	9: UMB/2−3	5.026(4.513)
4: DUN(1)-6	4.424(4.111)	9: UMB-5	4.868(4.544)	9: UMB/4−5	4.968(4.545)

4. An Unbalanced Incomplete Block Design Example

SCHEFFE (1959), p.189 gives the following data. Plates are washed with 5 detergent varieties in 10 blocks of size 3, in a balanced incomplete block design with the outcome given in Table 5. Obtain simultaneous 95% confidence intervals for

Table 5
Data from detergent experiment

Treatments	Blocks									
	1	2	3	4	5	6	7	8	9	10
1	27	28	30	31	29	30				
2	26	26	29				30	21	26	
3	30			34	32		34	31		33
4		29		33		34	31		33	31
5			26		24	25		23	24	26

all pairwise comparisons. (We shall assume the data for treatment 5 in blocks 9 and 10 are missing.)

Using the methodology described in section 2, we obtain the estimates of the treatment means (adjusted for block effects) as $t_1 = 27.1743$, $t_2 = 26.9646$, $t_3 = 30.8611$, $t_4 = 29.385$, $t_5 = 22.820$. The variance covariance matrix of the treatment mean estimates is

0.469006	0.265497	0.265497	0.328655	0.349708
0.265497	0.470700	0.263803	0.335673	0.325146
0.265497	0.263803	0.470700	0.335673	0.325146
0.328655	0.335673	0.335673	0.609357	0.367251
0.349708	0.325146	0.325146	0.367251	0.714620

The variance estimate s^2 is 2.0136 with 14 df. The q value is 4.40061. The 95% simultaneous confidence interval estimates for treatment i - treatment j are given in Table 6.

Table 6
95% Simultaneous All Pairs Confidence Intervals

(i, j)	2	3	4	5
1	(−1.595, 2.014)	(−5.491, −1.882)	(−4.070, -0.352)	(2.216, 6.492)
2		(−5.724, −2.069)	(−4.225, -0.616)	(1.782, 6.507)
3			(−0.329, 3.281)	(5.679, 10.404)
4				(3.962, 9.168)

5. Accuracies and Running Times

Accuracy and running times are functions mainly of the amount of Monte Carlo. For a specified accuracy running time is mainly a funtion of k and the number of simultaneous comparisons. Running time for an Athlon 800 processor with a Lahey LF95 Fortran compiler, for the Tukey procedure, with $k = 10$, $\alpha = 0.05$, is

approximately 0.4 seconds for an accuracy of $\pm.01$ (3 times the standard error) and approximately 17 seconds for an accuracy of $\pm.001$. Interactive calculations are quite practical. The SAS/V8.0 running times are approximately 2 and 90 minutes, respectively with a Celeron 333 MHz.

6. Summary and Conclusions

Efficient Fortran 90 and SAS-IML programs are available for obtaining critical values for multiple testing and simultaneous confidence intervals for 15 specific procedures and arbitrary sets of contrasts. The programs allow for missing values and covariates. The source code of the programs is available under the URL http://pegasus.cc.ucf.edu/~somervil/

References

BOFINGER, E., 1985: Multiple Comparisons and Type III errors. *J. Amer. Stat. Assoc.* **80**, 433–437.

BRETZ, F., 1999: Powerful Modifications of Williams' Test on Trend. Ph.D. dissertation, University of Hannover.

DUNNETT, C. W., 1955: A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc.* **50**, 1096–1121.

DUNNETT, C. W., 1964: New tables for multiple comparisons with a control. *Biometrics* **20**, 482–491.

HAYTER, A. J., 1990: A one-sided studentized range test for testing against a simple ordered alternative. *J. Amer. Stat. Assoc.* **85**, 778–785.

HIROTSU, C., 1997: Isotonic inference with particular interest in application to clinical trials. In: *Industrial Statistics*, KITSOS, C. P. and EDLER, L. (Eds.) Physica-Verlag, Heidelberg.

HSU, J. C., 1984: Constrained simultaneous confidence intervals for multiple comparisons with the best. *Ann. Stat.* **12**, 1136–1144.

LIU, W., MIWA, T., and HAYTER, A. J., 1999: Simultaneous confidence interval estimation for successive comparisons of ordered treatment effects. *J. Stat. Planning and Inf.* **88**, 75–86.

MC DERMOTT, M.P. and MUDHOLKAR, G. S., 1993: A simple approach to testing homogeneity or order-constrained means. *J. Amer. Stat. Assoc.* **88**, 1371–1379.

SCHEFFE, H., 1959: *The Analysis of Variance*. Wiley, N.Y.

SOMERVILLE, P. N., 1997: Multiple testing and simultaneous confidence intervals: calculation of constants. *Computational Statistics and Data Analysis* **25**, 217–223.

SOMERVILLE, P. N., 1999: Critical values for multiple testing and comparisons – one step and stepdown procedures. *Journal of Statistical Planning and Inference* **82**, 129–138.

SOMERVILLE, P. N., LIU, W., MIWA, T., and HAYTER, A. J., 2001: Combining one-sided and two-sided confidence interval procedures for successive comparisons of ordered treatment effects. Biometrical Journal **43**, 5, 533–542.

TUKEY, J. W., 1953: The problem of multiple comparisons. Mimeographed Monograph (1953).

Dr. PAUL SOMERVILLE
Department of Statistics
University of Central Florida
Orlando, FL 32816–2370
E-mail: somervil@pegasus.cc.ucf.edu

Received, September 2000
Revised, April 2001
Accepted, May 2001