Multiple Comparison Procedures under Heteroscedasticity

Shun-Yi Chen* and Su-Hao Lee

Department of Mathematics, Tamkang University, Tamsui, Taiwan 251, R.O.C.

Abstract

When the homogeneity hypothesis of treatment means under heteroscedasticity is rejected, one may be interested in comparing each treatment with all other treatments. The objective of this paper is to derive two multiple range test procedures, the Newman-Keuls type and Duncan type, for multiple comparisons of treatment means with unknown and unequal variances. We investigate these pairwise comparison procedures with the Tukey type simultaneous confidence intervals approach with respect to comparisonwise and experimentwise error rates, as well as correct decision rates. The tables for applying the new multiple range test procedures are provided, along with a numerical example to demonstrate the methods.

Key Words: Pairwise Comparisons, One-Stage and Two-Stage Sampling Procedures, Significance Level, Protection Level, Correct Decision Rate

1. Introduction

Suppose that I (≥ 2) independent populations $\pi_1, \ldots, \pi_I$ are available, where observations taken from population π_i are normally distributed with mean μ_i and variance σ_i^2 ($1 \leq i \leq I$). It is assumed that these variances are unknown and unequal, and no prior knowledge about μ_i and σ_i^2 is available.

The procedures for the conventional analysis of variance (ANOVA) are based on the assumption that the error variances are equal. When an ANOVA test for testing equal treatment means is rejected, many multiple comparison procedures are available for detecting the specific differences between treatment means, such as the least significant difference (LSD) method, Tukey's test, Scheffé's method, Newman-Keuls' procedure, and Duncan's multiple range test, etc. With the criteria of correct decision rate and error rates, it is found that, by using Monte Carlo simulations, Scheffé's, Tukey's and Newman-Keuls' tests are less appropriate than LSD, Waller-Duncan method, or Duncan's multiple range test [1].

In the situation where the variances are unknown and unequal, an analysis of variance procedure for the means of I independent normal populations is developed by using a two-stage sampling procedure [2]. A two-stage range test procedure for testing the hypothesis of equal means in an analysis of variance problem is also derived [3]. Chen [4] provided one-stage and two-stage statistical inferences for testing the equality of means and showed that the two-stage range test and the two-stage ANOVA test are equivalent, because they produce the same sample sizes at given levels and powers. As a follow-up to the test, Scheffé type or Tukey type multiple comparison procedures have been considered [5,6]. Tamhane [7] also proposed exact two-stage procedures for multiple comparison problems. Wilcox [8] provided a table of percentage points for the range of I independent t random variables which can be applied to the above multiple comparison procedures.

Our goal in this article is to derive Newman-Keuls type and Duncan type multiple range tests for the multiple comparison procedure under heteroscedasticity, and to provide extensive tables of the percentage points to implement the new procedures. In section 2 we introduce

*Corresponding author. E-mail: sychen@mail.tku.edu.tw

the two-stage sampling procedure and the tests for testing the equality of means in the analysis of variance problem. Section 3 presents two-stage multiple comparison procedures and the corresponding tables of percentage points. In section 4, we propose the one-stage sampling procedure and the one-stage multiple comparison procedures. In section 5, we investigate three multiple comparison procedures under unequal variances using simulation studies. To illustrate the use of the procedures, we provide a numerical example in section 6.

2. The Two-Stage Sampling Procedure

The model we are considering for the one-way layout is:

$$X_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1,\ldots,I, \quad j = 1,2,\ldots \tag{1}$$

where the e_{ij}'s are independent random variables with $e_{ij} \sim N(0,\sigma_i^2)$ and $\sum_{i=1}^{I}\alpha_i = 0$. We may denote mean μ_i by $\mu_i = \mu + \alpha_i$. The variances σ_i^2's are unknown and possibly unequal. The two-stage sampling procedure (**P_2**) for testing the equality of means is given as follows [2]:

P_2: Choose a number $z > 0$ (z is determined by the power of the test), and take an initial sample of size n_0 (at least 2, but 10 or more will give better results) from each of the I populations. For the ith population let S_i^2 be the usual unbiased estimate of σ_i^2 based on the first n_0 observations, and define

$$N_i = \max\left\{n_0 + 1, \left[\frac{S_i^2}{z}\right] + 1\right\} \tag{2}$$

where $[x]$ denotes the greatest integer less than or equal to x. Then, take $N_i - n_0$ additional observations for the second stage from the ith population so that we have a total of N_i observations denoted by $X_{i1}, \ldots, X_{in_0}, \ldots, X_{iN_i}$. For each i, set the coefficients $a_{i1}, \ldots, a_{in_0}, \ldots, a_{iN_i}$, so that

$$a_{i1} = \cdots = a_{in_0} = \frac{1-(N_i - n_0)b_i}{n_0} = a_i,$$

$$a_{i,n_0+1} = \cdots = a_{iN_i} = \frac{1}{N_i}\left[1 + \sqrt{\frac{n_0(N_i z - S_i^2}{(N_i - n_0)S_i^2}}\right] = b_i,$$

and then compute the weighted mean

$$\tilde{X}_{i.} = a_i \sum_{j=1}^{n_0} X_{ij} + b_i \sum_{j=n_0+1}^{N_i} X_{ij} \tag{3}$$

which is a linear combination of the first-stage data (X_{i1}, ..., X_{in_0}) and the second-stage data ($X_{i,n_0+1}, \ldots, X_{iN_i}$). It was proved that the random variables

$$t_i = \frac{\tilde{X}_{i.} - \mu_i}{\sqrt{z}}, \quad i = 1,\ldots,I,$$

have i.i.d. t distribution with $n_0 - 1$ degrees of freedom [9]. The distribution of $\tilde{X}_{i.}$ is independent of the unknown variances σ_i^2.

The ANOVA test statistic given by Bishop and Dudewicz [2] for testing the null hypothesis H_0:$\mu_1 = \cdots = \mu_I$ against the alternative $H_a: \sum(\mu_i - \bar{\mu})^2 \geq \delta > 0$ is

$$F = \sum_{i=1}^{I} \frac{(\tilde{X}_{i.} - \tilde{X}_{..})^2}{z} \tag{4}$$

where $\bar{\mu} = \sum \mu_i / I$ and $\tilde{X}_{..}$ is the mean of $\tilde{X}_{1.}, \ldots, \tilde{X}_{I.}$. The null hypothesis H_0 is rejected at level α if and only if $F > F_\alpha$. The upper α percentage point F_α and tables of power- and δ-related z values can be found in [2].

Chen and Chen [3] proposed a two-stage range test to test H_0:$\mu_1 = \cdots = \mu_I$ against H_a:$\mu_{max} - \mu_{min} \geq \delta^* > 0$, where μ_{max}(μ_{min}) is the maximum(minimum) of μ_1, …, μ_I. The range test statistic is given by

$$R = \frac{\tilde{X}_{max} - \tilde{X}_{min}}{\sqrt{z}} \tag{5}$$

where $\tilde{X}_{max}$ ($\tilde{X}_{min}$) is the maximum(minimum) of $\tilde{X}_{1.}$, ..., $\tilde{X}_{I.}$. The null hypothesis of $\mu_1 = \cdots = \mu_I$ is rejected at level α if and only if $R > R_\alpha$, where R_α is the level α critical value and z is determined by the power of the test P^* such that

$$\inf P(R > R_\alpha \mid H_a : \mu_{max} - \mu_{min} \geq \delta^* > 0) = P^*$$

where $\delta^* > 0$ is specified in such a way that both the level and the power of the test are controllable. The power- and δ^*-related z values can be obtained from tables given by [3].

When the homogeneity hypothesis of treatment means under heteroscedasticity is rejected, one may be interested in comparing each treatment with all other treatments. Following the above tests of equality of means, we present, in the next section, two multiple range test procedures for comparing all pairs of I treatment means: the two-stage Newman-Keuls type (NK2) and two-stage Duncan type (D2).

3. Multiple Comparison Procedures

Assume that for $i = 1, \ldots, I$, the two-stage sampling procedure ($\mathbf{P_2}$) has been conducted, and that the final weighted sample means $\tilde{X}_{i.}$'s have been computed as in (3). We rank these weighted sample means in ascending order as $\tilde{X}_{(1)}, \tilde{X}_{(2)}, \ldots, \tilde{X}_{(I)}$, where $\tilde{X}_{(1)}$ denotes the smallest weighted mean, and $\tilde{X}_{(I)}$ denotes the largest weighted mean. The null hypotheses that we wish to test are $H_0{:}\mu_i = \mu_j$ for all $i \neq j$ at α significance level. The test statistic of the two-stage Newman-Keuls type (NK2) procedure is given by

$$K_p = q_{\alpha,p,\nu}\sqrt{z},\ p = 2,3,\ldots,I \tag{6}$$

where $q_{\alpha,p,\nu}$ is the upper percentage point of the range of p independent t random variables each with $\nu = n_0 - 1$ degrees of freedom, and z-value is defined in the two-stage sampling procedure ($\mathbf{P_2}$). There are $I(I-1)/2$ pairs of mean differences to be tested, starting with the difference between the largest weighted mean and the smallest, $\tilde{X}_{(I)} - \tilde{X}_{(1)}$, which is then compared with K_I. Next we examine the difference between the largest weighted mean and the second smallest, $\tilde{X}_{(I)} - \tilde{X}_{(2)}$, which is compared with K_{I-1}. The comparison procedures are continued until all possible pairs of means have been considered, i.e., until the difference between the second smallest and the smallest mean, $\tilde{X}_{(2)} - \tilde{X}_{(1)}$, is compared with K_2. In general, for the comparison of the means of the pair (k, m), where $1 \le k < m \le I$ and $p = m - k + 1$, if

$$\tilde{X}_{(m)} - \tilde{X}_{(k)} > K_p \tag{7}$$

then the means of the pair (k, m) are concluded significantly different; otherwise this pair of means and all the means that fall between, i.e. $\{\tilde{X}_{(j)}, k \le j \le m\}$, are not significantly different and will be included in a single group. For the means included in the same group, no further tests will be performed. That is, no subset of means grouped in a nonsignificant group can later be considered significant. Under H_0, the upper percentage point $q_{\alpha,p,\nu}$ of the range of p independent t random variables can be obtained from

$$\begin{aligned}
\alpha &= P(\tilde{X}_{(m)} - \tilde{X}_{(k)} > q_{\alpha,p,\nu}\sqrt{z}\,|\,H_0) \\
&= 1 - P(\tilde{X}_{(m)} \le \tilde{X}_{(k)} + q_{\alpha,p,\nu}\sqrt{z}\,|\,H_0) \\
&= 1 - \sum_{j=k}^{m} P(\tilde{X}_{(j)} < \tilde{X}_{(i)} \le \tilde{X}_{(j)} + q_{\alpha,p,\nu}\sqrt{z},\ k \le i \le m,\ i \ne j\,|\,H_0) \\
&= 1 - \sum_{j=k}^{m} P(T_j < T_i \le T_j + q_{\alpha,p,\nu},\ k \le i \le m,\ i \ne j) \\
&= 1 - p\int_{-\infty}^{\infty} [F(y + q_{\alpha,p,\nu}) - F(y)]^{p-1} f(y)dy
\end{aligned} \tag{8}$$

where $F(\cdot)$ and $f(\cdot)$ are the cumulative distribution function and probability density function of random variable T_i which has t distribution with ν degrees of freedom. The tables of percentage points $q_{\alpha,p,\nu}$ can be found in [3] and [8].

The two-stage Duncan type multiple range test procedure (D2) is derived by applying the protection level $(1-\alpha)^{p-1}$ proposed by Duncan [10]. The D2 comparison procedure is the same as that for NK2, except that the upper percentage points are calculated differently. In other words, the mean differences $\tilde{X}_{(m)} - \tilde{X}_{(k)}$ will now be compared with D_p, and

$$D_p = \gamma_{\alpha,p,\nu}\sqrt{z},\quad p = 2,3,\ldots,I \tag{9}$$

where $\gamma_{\alpha,p,\nu}$ is the upper percentage point of the range of p independent t random variables each with $\nu = n_0 - 1$ degrees of freedom at the error rate of $\alpha_p = 1 - (1 - \alpha)^{p-1}$. The values of $\gamma = \gamma_{\alpha,p,\nu}$ can be obtained from

$$g(\gamma) = P(\gamma) - \alpha_p = 0 \tag{10}$$

where

$$\begin{aligned}
P(\gamma) &= P(\tilde{X}_{(m)} - \tilde{X}_{(k)} > \gamma\sqrt{z}\,|\,H_0) \\
&= 1 - p\int_{-\infty}^{\infty} [F(y+\gamma) - F(y)]^{p-1} f(y)dy
\end{aligned} \tag{11}$$

Using the Newton-Raphson iteration method, we have, at the nth iteration, $n = 1, 2, \ldots$,

$$\gamma^{(n)} = \gamma^{(n-1)} - [\partial g(\gamma)/\partial\gamma]^{-1} \cdot g(\gamma^{(n-1)}) \tag{12}$$

where $\partial g(\gamma)/\partial\gamma$ is the partial derivative evaluated at the $(n-1)$th iteration, and

$$\partial g(\gamma)/\partial\gamma = -p(p-1)\int_{-\infty}^{\infty}[F(y+\gamma)-F(y)]^{p-2}f(y+\gamma)f(y)dy$$

The values of $\gamma = \gamma_{\alpha,p,\nu}$ are calculated using the same numerical method as described in [3], and they are reported in Tables 1–3 for α = .10, .05, .01, p = 2(1)10(2)20, 50, 100 and ν = 1(1)20, 24, 29, 39, 59, 99, ∞.

The two-stage Tukey type (T2) simultaneous confidence intervals [5,8], for multiple comparisons of treatment means under unequal variances, are similar to that of NK2 and D2, except that the test statistic of T2 at α level is

$$T_p = q_{\alpha,I,\nu}\sqrt{z}, \quad p = 2,3,...,I \tag{13}$$

where $q_{\alpha,I,\nu}$ is the upper percentage point of the range of I independent t random variables each with $\nu = n_0 - 1$ degrees of freedom. The upper percentage point $q_{\alpha,I,\nu}$ can be found in [3] and [8].

4. The One-Stage Multiple Comparison Procedures

The two-stage procedure discussed in section 2 is a design-oriented statistical method which depends on the final sample size N_i in (2) to meet a prespecified power and level. However, in situations where the experiment is terminated early due to unexpected reasons, the required sample size N_i in the two-stage procedure cannot be achieved. In this situation we have to employ the available n_i, $i = 1, \ldots, I$, observations on hand to recalculate the coefficients in (3) using the one-stage sampling procedure [11]. The one-stage procedure ($\mathbf{P_1}$) is briefly described below.

$\mathbf{P_1}$: Given the model in (1), use the first (or any randomly chosen) n_0 ($2 \le n_0 < n_i$) observations from each of the I populations to calculate the usual unbiased sample variance, denoted by S_i^2. Calculate weights U_i and V_i for the observations in the ith sample as

$$U_i = \frac{1}{n_i} + \frac{1}{n_i}\sqrt{\frac{n_i - n_0}{n_0}(n_i z^* / S_i^2 - 1)}$$

$$V_i = \frac{1}{n_i} - \frac{1}{n_i}\sqrt{\frac{n_0}{n_i - n_0}(n_i z^* / S_i^2 - 1)}$$

where z^* is the maximum of $\{S_i^2 / n_i,\ i = 1, \ldots, I\}$. Compute the final weighted sample mean using all observations by

$$\tilde{X}_{i.} = \sum_{j=1}^{n_0} U_i X_{ij} + \sum_{j=n_0+1}^{n_i} V_i X_{ij} \tag{14}$$

It is known that the random variables

$$t_i = \frac{\tilde{X}_{i.} - \mu_i}{\sqrt{z^*}}, \quad i = 1, \ldots, I,$$

are distributed as independent Student's t with $\nu = n_0 - 1$ degrees of freedom [9]. Chen [4] proposed the one-stage range test and one-stage ANOVA test for testing the equality of treatment means. Note that given the sample data, power can also be determined; it is data-dependent and can be larger than, equal to, or smaller than the originally specified one [4]. Similar to the procedures in section 3, we propose the following one-stage multiple range test procedures for pairwise comparisons: the one-stage Newman-Keuls type (NK1), one-stage Duncan type (D1) and one-stage Tukey type (T1).

Assume the one-stage sampling procedure ($\mathbf{P_1}$) has been conducted and the weighted sample means $\tilde{X}_{i.}$'s have been computed as in (14). The weighted sample means are ranked in ascending order as $\tilde{X}_{(1)}$, $\tilde{X}_{(2)}$, ..., $\tilde{X}_{(I)}$. The NK1 procedure is the same as NK2 except that the test statistic of NK1 is given by

$$K_p = q_{\alpha,p,\nu}\sqrt{z^*}, \quad p = 2,3,...,I \tag{15}$$

where the upper percentage point $q_{\alpha,p,\nu}$ is defined in (8) and can be found in [3] and [8].

The one-stage Duncan type multiple range test procedure (D1) is also the same as D2 except that the test statistic of D1 at α level is given by

$$D_p = \gamma_{\alpha,p,\nu}\sqrt{z^*}, \quad p = 2,3,...,I \tag{16}$$

where $\gamma_{\alpha,p,\nu}$ is the upper percentage point defined in (9) and can be obtained from Tables 1–3. Similarly, the test statistic of one-stage Tukey type (T1) simultaneous con-

Table 1. Upper percentage points $\gamma_{\alpha,p,\nu}$ of the two-stage Duncan type multiple comparison (D2) for α = 0.10

	p								
ν	2	3	4	5	6	7	8	9	10
1	12.63	10.64	10.07	9.87	9.79	9.78	9.79	9.82	9.86
2	4.57	4.49	4.51	4.56	4.61	4.65	4.70	4.73	4.77
3	3.50	3.56	3.63	3.69	3.75	3.80	3.84	3.87	3.91
4	3.11	3.21	3.29	3.36	3.41	3.46	3.50	3.53	3.56
5	2.91	3.03	3.11	3.18	3.23	3.28	3.32	3.35	3.38
6	2.79	2.91	3.00	3.07	3.13	3.17	3.21	3.24	3.27
7	2.71	2.84	2.93	3.00	3.05	3.10	3.13	3.17	3.19
8	2.66	2.79	2.88	2.94	3.00	3.04	3.24	3.11	3.14
9	2.62	2.75	2.84	2.90	2.96	3.00	3.04	3.07	3.10
10	2.58	2.71	2.80	2.87	2.93	2.97	3.01	3.04	3.07
12	2.53	2.67	2.76	2.83	2.88	2.92	2.96	2.99	3.02
14	2.50	2.64	2.73	2.80	2.85	2.89	2.93	2.96	2.99
19	2.45	2.59	2.68	2.75	2.80	2.84	2.88	2.91	2.94
24	2.42	2.56	2.65	2.72	2.77	2.81	2.85	2.88	2.91
29	2.40	2.54	2.63	2.70	2.75	2.79	2.83	2.86	2.88
39	2.38	2.52	2.61	2.68	2.73	2.77	2.81	2.84	2.86
59	2.36	2.50	2.59	2.66	2.71	2.75	2.79	2.82	2.84
99	2.35	2.48	2.58	2.64	2.69	2.74	2.77	2.80	2.83
∞	2.33	2.46	2.55	2.62	2.67	2.71	2.75	2.78	2.80
ν	12	14	16	18	20	50	100		
1	9.95	10.04	10.12	10.21	10.28	10.98	12.92		
2	4.83	4.89	4.94	4.98	5.02	5.31	5.85		
3	3.96	4.01	4.05	4.08	4.11	4.33	4.70		
4	3.62	3.66	3.69	3.73	3.75	3.95	4.25		
5	3.43	3.47	3.51	3.54	3.56	3.74	4.02		
6	3.32	3.36	3.39	3.42	3.44	3.62	3.87		
7	3.24	3.28	3.31	3.34	3.36	3.53	3.77		
8	3.19	3.22	3.26	3.28	3.31	3.47	3.70		
9	3.14	3.18	3.21	3.24	3.26	3.42	3.65		
10	3.11	3.15	3.18	3.21	3.23	3.39	3.61		
12	3.06	3.10	3.13	3.16	3.18	3.33	3.54		
14	3.03	3.07	3.10	3.12	3.14	3.30	3.50		
19	2.98	3.01	3.04	3.07	3.09	3.24	3.44		
24	2.95	2.98	3.01	3.04	3.06	3.21	3.40		
29	2.93	2.96	2.99	3.02	3.04	3.18	3.38		
39	2.91	2.94	2.97	2.99	3.01	3.16	3.35		
59	2.89	2.92	2.95	2.97	2.99	3.14	3.32		
99	2.87	2.90	2.93	2.95	2.97	3.12	3.30		
∞	2.84	2.88	2.91	2.93	2.95	3.09	3.27		

fidence intervals for multiple comparisons at α level is

$$T_p = q_{\alpha,I,\nu}\sqrt{z^*}, \quad p = 2,3,\ldots,I \tag{17}$$

where the upper percentage point $q_{\alpha,I,\nu}$ is defined in (13) and can be found in [3] and [8].

5. Numerical Studies

In this section, we first investigate three two-stage multiple comparison procedures, NK2, D2 and T2, with respect to the comparisonwise error rate and experimentwise error rate by using the Monte Carlo simulation. The comparisonwise and experimentwise error rates are defined in [1]. The comparisonwise error rate refers to the ratio of the number of type I error made to the total number of inferences conducted in all experiments analyzed. The experimentwise error rate is the ratio of the number of experiments in which at least one type I error is committed divided by the total number of experiments

Table 2. Upper percentage points $\gamma_{\alpha,p,\nu}$ of the two-stage Duncan type multiple comparison (D2) for $\alpha = 0.05$

	p								
ν	2	3	4	5	6	7	8	9	10
1	25.41	20.64	19.21	18.64	18.39	18.29	18.26	18.28	18.33
2	6.54	6.33	6.33	6.38	6.44	6.50	6.55	6.61	6.66
3	4.59	4.64	4.72	4.80	4.89	4.93	4.98	5.03	5.07
4	3.94	4.05	4.15	4.23	4.30	4.36	4.41	4.45	4.49
5	3.63	3.76	3.86	3.94	4.01	4.06	4.11	4.15	4.19
6	3.45	3.59	3.69	3.77	3.84	3.89	3.94	3.98	4.01
7	3.33	3.47	3.57	3.65	3.72	3.77	3.82	3.86	3.89
8	3.24	3.39	3.49	3.57	3.64	3.69	3.73	3.77	3.81
9	3.18	3.33	3.43	3.51	3.57	3.63	3.67	3.71	3.74
10	3.13	3.28	3.39	3.46	3.53	3.58	3.62	3.66	3.69
12	3.06	3.21	3.32	3.39	3.46	3.51	3.55	3.59	3.62
14	3.02	3.17	3.27	3.35	3.41	3.46	3.50	3.54	3.57
19	2.95	3.10	3.20	3.27	3.34	3.38	3.43	3.46	3.49
24	2.91	3.06	3.16	3.23	3.29	3.34	3.38	3.42	3.45
29	2.88	3.03	3.13	3.20	3.26	3.31	3.35	3.39	3.42
39	2.85	3.00	3.10	3.17	3.23	3.28	3.32	3.36	3.39
59	2.82	2.97	3.07	3.15	3.20	3.25	3.29	3.32	3.35
99	2.80	2.95	3.05	3.12	3.18	3.23	3.27	3.30	3.33
∞	2.77	2.92	3.02	3.09	3.15	3.19	3.23	3.26	3.29
ν	12	14	16	18	20	50	100		
1	18.47	18.62	18.78	18.94	19.10	20.66	21.87		
2	6.75	6.84	6.91	6.97	7.03	7.52	7.84		
3	5.15	5.21	5.27	5.31	5.36	5.69	5.90		
4	4.56	4.61	4.66	4.70	4.74	5.02	5.19		
5	4.25	4.31	4.35	4.39	4.42	4.68	4.82		
6	4.07	4.12	4.16	4.20	4.23	4.47	4.61		
7	3.95	4.00	4.04	4.07	4.10	4.33	4.46		
8	3.86	3.91	3.95	3.98	4.01	4.23	4.36		
9	3.80	3.84	3.88	3.91	3.94	4.16	4.28		
10	3.75	3.79	3.83	3.86	3.89	4.10	4.22		
12	3.67	3.72	3.75	3.78	3.81	4.02	4.13		
14	3.62	3.66	3.70	3.73	3.76	3.96	4.07		
19	3.54	3.59	3.62	3.65	3.68	3.87	3.97		
24	3.50	3.54	3.58	3.61	3.63	3.82	3.92		
29	3.47	3.51	3.54	3.57	3.60	3.78	3.88		
39	3.44	3.48	3.51	3.54	3.56	3.75	3.84		
59	3.40	3.44	3.48	3.51	3.53	3.71	3.81		
99	3.38	3.42	3.45	3.48	3.50	3.68	3.78		
∞	3.34	3.38	3.41	3.44	3.47	3.61	3.67		

analyzed. For the one-way layout with four independent populations, the initial sample of size $n_0 = 10$ is chosen from each population. Using the table from [3], the z value should be 0.0265 so that the power of the range test is equal to 0.90 for $\delta^* = 1$ in the alternative hypothesis. The variances are (1,2,3,4), (1,4,4,9) and (1,1,1,9). Under every configuration of variances and homogeneity among population means, 10,000 simulation runs are performed. On each simulation run, observations are generated from the standard normal distribution using the random number generator RANNOR in SAS 8.2 [12], and three two-stage multiple comparison procedures are conducted. The error rates for each configuration are calculated after the 10,000 runs and are listed in Table 4. It can be seen from the table that, in general, the comparisonwise error rates of the D2 are close to the nominal level α; while the experimentwise error rates for the NK2 and T2 are close to the nominal level α. The D2 also shows close agreement between observed and expected experimentwise error rates α_p; where $\alpha_p = 1 - (1 -$

Table 3. Upper percentage points $\gamma_{\alpha,p,\nu}$ of the two-stage Duncan type multiple comparison (D2) for α = 0.01

	p								
ν	2	3	4	5	6	7	8	9	10
1	127.55	98.26	88.62	84.35	82.02	80.76	79.94	79.33	79.08
2	14.43	13.41	13.18	13.14	13.17	13.23	13.30	13.37	13.45
3	7.92	7.85	7.91	8.00	8.08	8.17	8.24	8.31	8.39
4	6.15	6.25	6.36	6.46	6.55	6.63	6.70	6.76	6.82
5	5.39	5.53	5.66	5.76	5.85	5.92	5.99	6.04	6.09
6	4.97	5.13	5.26	5.36	5.44	5.51	5.58	5.63	5.68
7	4.71	4.88	5.01	5.11	5.19	5.25	5.31	5.36	5.41
8	4.53	4.71	4.83	4.93	5.01	5.07	5.13	5.18	5.22
9	4.41	4.58	4.70	4.80	4.88	4.94	5.00	5.04	5.09
10	4.31	4.49	4.61	4.70	4.78	4.86	4.89	4.94	4.98
12	4.18	4.35	4.47	4.56	4.63	4.70	4.75	4.79	4.83
14	4.09	4.26	4.38	4.47	4.54	4.60	4.65	4.69	4.73
19	3.96	4.12	4.24	4.33	4.39	4.45	4.50	4.54	4.58
24	3.89	4.05	4.16	4.25	4.32	4.37	4.42	4.46	4.50
29	3.83	4.00	4.11	4.19	4.26	4.31	4.36	4.40	4.44
39	3.78	3.94	4.05	4.13	4.20	4.25	4.30	4.34	4.38
59	3.73	3.89	4.00	4.08	4.15	4.20	4.24	4.28	4.32
99	3.70	3.85	3.96	4.04	4.10	4.15	4.20	4.24	4.27
∞	3.64	3.80	3.90	3.98	4.04	4.09	4.14	4.17	4.20
ν	12	14	16	18	20	50	100		
1	78.91	79.03	79.33	79.63	80.06	85.95	92.40		
2	13.61	13.75	13.90	14.02	14.14	15.30	16.22		
3	8.50	8.60	8.70	8.78	8.85	9.52	10.00		
4	6.92	7.01	7.08	7.15	7.21	7.71	8.07		
5	6.18	6.26	6.33	6.38	6.65	6.86	7.16		
6	5.76	5.83	5.89	5.95	5.99	6.38	6.63		
7	5.49	5.56	5.61	5.66	5.71	6.06	6.30		
8	5.30	5.36	5.42	5.47	5.51	5.84	6.06		
9	5.16	5.22	5.28	5.32	5.36	5.68	5.89		
10	5.05	5.11	5.17	5.21	5.25	5.56	5.76		
12	4.90	4.96	5.01	5.05	5.09	5.38	5.57		
14	4.80	4.85	4.90	4.94	4.98	5.26	5.44		
19	4.64	4.70	4.74	4.78	4.81	5.08	5.25		
24	4.56	4.61	4.65	4.69	4.72	4.98	5.14		
29	4.50	4.55	4.59	4.62	4.66	4.91	5.07		
39	4.43	4.48	4.53	4.56	4.59	4.84	4.99		
59	4.37	4.42	4.46	4.50	4.53	4.77	4.92		
99	4.33	4.38	4.42	4.45	4.48	4.71	4.86		
∞	4.26	4.31	4.34	4.38	4.41	4.60	4.68		

$\alpha)^{p-1}$ = .27, .14 and .03 for α = .10, .05 and .01, respectively. The error rates of the NK2 are slightly higher than that of the T2 because the critical value of the NK2 changes as p changes, for $p = 2, 3, \ldots, I$, which is smaller than the critical value of the T2 when $p < I$, and consequently the NK2 would generate higher error rates than the T2. The simulation results suggest that, for the multiple comparison procedure under heteroscedasticity, if statisticians are concerned with protecting the comparisonwise Type I error rate then they may choose the D2 procedure. If statisticians want to control the experimentwise type I error rate, they may use the NK2 or T2 procedure.

We next examine the ability to correctly detect real differences between means, comparisonwise correct decision rate and experimentwise correct decision rate, for the three multiple comparison procedures when α = .05. The simulation process is the same as that described above. The variance configurations are (1,2,3,4), (1,4,4,9) and (1,1,1,9). The correct decision rates for 10,000

Table 4. The comparisonwise and experimentwise error rates of D2, NK2 and T2 procedures

variances	comparisonwise			experimentwise		
	D2	NK2	T2	D2	NK2	T2
			$\alpha = .10$			
(1,2,3,4)	.079	.048	.033	.254	.138	.120
(1,4,4,9)	.080	.049	.034	.252	.141	.118
(1,1,1,9)	.101	.067	.048	.297	.180	.159
			$\alpha = .05$			
(1,2,3,4)	.041	.020	.015	.135	.059	.052
(1,4,4,9)	.041	.022	.016	.141	.063	.055
(1,1,1,9)	.059	.036	.028	.180	.095	.086
			$\alpha = .01$			
(1,2,3,4)	.010	.006	.005	.033	.018	.016
(1,4,4,9)	.009	.006	.005	.032	.017	.016
(1,1,1,9)	.019	.014	.012	.057	.034	.033

† D2 = Two-stage Duncan type procedure.
NK2 = Two-stage Newman-Keuls type procedure.
T2 = Two-stage Tukey type procedure.

simulation runs under various combinations of population means are listed in Table 5. Table 5 shows that, as expected, the D2 procedure produces larger comparisonwise and experimentwise correct decision rates than that of the NK2 and T2. The discrepancy of correct decision rates across the D2 versus NK2 and T2 are due to the different types of error rates they are controlling. From Table 4 we can see that the T2, which is controlling the experimentwise type I error rate, is more conservative than either the NK2 or D2; and consequently the T2 has less correct decision rate than either the NK2 or D2 procedure. We have also compared the correct decision rates for other values of I and α, and found similar results to those reported in Table 5.

6. A Numerical Example

Bishop and Dudewicz [2] studied the data of the bacterial killing ability of four solvents using the two-stage sampling procedure. There were four types of solvents that could have affected the ability of the fungicide methyl-2-benzimidazole-carbamate to destroy the fungus Penicillium expansum. The fungicide was diluted in exactly the same manner in four different types of solvents and sprayed on the fungus. The percentage of fungus destroyed was measured and recorded. In the first stage of their experiment, $n_0 = 15$ observations were run with each solvent. A robust Levene's test was performed [13] for the homogeneity of variances and found a significant difference among the variances (p-value < .001). The two-stage sampling procedure ($\mathbf{P_2}$) had been conducted where $z = .08$ was chosen at $\alpha = .10$ to achieve the power of .85. The final sample sizes (N_i) required for each solvent were $N_1 = 27$, $N_2 = 40$, $N_3 = 74$ and $N_4 = 16$. The final weighted sample means were found to be $\tilde{X}_{1.} = 97.55$, $\tilde{X}_{2.} = 95.25$, $\tilde{X}_{3.} = 94.75$ and $\tilde{X}_{4.} = 97.49$. The null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$ was rejected by using the ANOVA test in (4) and the range test in (5). Following the tests, to perform the two-stage multiple comparison procedures, the weighted means were ranked from the smallest to the largest, and the differences of the means for each pair were computed. Table 2 shows that $\gamma_{\alpha, p, \nu}$, the upper percentage points of D2, with $\nu = n_0 - 1 = 14$ degrees of freedom are 3.27, 3.17, and 3.02 for p = 4, 3, and 2, respectively. The upper percentage points of NK2 can be found from Wilcox [8]: $q_{\alpha,p,\nu}$ = 4.03, 3.65, and 3.02 for p = 4, 3, and 2. The pairwise differences and the corresponding test statistics are summarized in Table 6. We conclude that solvent 1 and solvent 4 are not significantly different, and both are in the same group (Group 1); while solvent 2 is not significantly different from solvent 3 and both are in the same group (Group 2). The solvents in Group 1 are superior to those in Group 2. Note that we have obtained the same conclusion as Bishop [5] and Wilcox [8] by using D2 and

Table 5. The comparisonwise and experimentwise correct decision rates of D2, NK2 and T2 procedures when α = .05

	comparisonwise			experimentwise		
mean	D2	NK2	T2	D2	NK2	T2
			variance: (1,2,3,4)			
(1,1,1,1)	.041	.021	.015	.138	.061	.054
(1,1.1,1.2,1.3)	.060	.028	.021	.208	.089	.078
(1,1.2,1.4,1.6)	.171	.102	.071	.521	.301	.274
(1,1.3,1.6,1.9)	.352	.273	.197	.848	.687	.654
(1,1.4,1.8,2.2)	.523	.467	.356	.978	.925	.916
(1,1.5,2,2.5)	.644	.616	.488	.997	.989	.985
(1,2,3,4)	.928	.927	.850	1	1	1
			variance: (1,4,4,9)			
(1,1,1,1)	.039	.020	.015	.131	.059	.051
(1,1.1,1.2,1.3)	.056	.025	.019	.201	.082	.078
(1,1.2,1.4,1.6)	.162	.093	.065	.498	.283	.259
(1,1.3,1.6,1.9)	.341	.263	.191	.826	.661	.633
(1,1.4,1.8,2.2)	.512	.457	.343	.968	.910	.897
(1,1.5,2,2.5)	.637	.610	.482	.996	.987	.984
(1,2,3,4)	.921	.920	.839	1	1	1
			variance: (1,1,1,9)			
(1,1,1,1)	.058	.036	.029	.178	.095	.086
(1,1.1,1.2,1.3)	.077	.046	.035	.242	.130	.119
(1,1.2,1.4,1.6)	.184	.118	.085	.540	.339	.313
(1,1.3,1.6,1.9)	.369	.292	.215	.850	.699	.671
(1,1.4,1.8,2.2)	.532	.483	.371	.978	.923	.912
(1,1.5,2,2.5)	.645	.620	.496	.996	.988	.985
(1,2,3,4)	.904	.903	.840	1	1	1

NK2 procedures, but the values of the test statistics of D2 and NK2 are smaller than that of T2.

If the experiment at the second stage is terminated earlier due to some uncontrollable factors, and we can only obtain smaller samples from each treatment, say, N_1 = 24, N_2 = 36, N_3 = 67, and N_4 = 16. We can apply the one-stage procedure in section 4 with those observations available, i.e. first N_i (i = 1, ..., 4) observations, from each population. For this case we use the same number of observations used by Bishop and Dudewicz [2], n_0 = 15 for initial estimation and the remaining for use in the final computation of the one-stage procedure. The intermediate statistics S_i^2, U_i, V_i, and z^* are given in Table 7. The final weighted sample means are $\widetilde{X}_1 = 97.43$, $\widetilde{X}_2 = 94.90$, $\widetilde{X}_3 = 94.86$, and $\widetilde{X}_4 = 97.23$. The null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$ was rejected by using the one-stage ANOVA test and the one-stage range test [4]. For the one-stage multiple comparison procedures, the weighted means were ranked from the smallest to the largest and the differences of the means for each pair were computed. Since the initial sample size is n_0 = 15 for the one-stage procedure, we have $\nu = n_0 - 1 = 14$ degrees of freedom; the percentage points of D1 and NK1 are the same as those used for D2 and NK2. The pairwise differences and the corresponding test statistics are summarized in Table 8. We obtain the same conclusion as that in the two-stage procedure, i.e., solvent 1 and solvent 4 are in the same

Table 6. The two-stage multiple comparison procedures for bacterial killing ability example

	test statistic		
pairwise difference	D2	NK2	T2
$\tilde{X}_{1.} - \tilde{X}_{3.} = 2.80$	0.92	1.14	1.14
$\tilde{X}_{1.} - \tilde{X}_{2.} = 2.30$	0.90	1.03	1.14
$\tilde{X}_{1.} - \tilde{X}_{4.} = 0.06$	0.85	0.85	1.14
$\tilde{X}_{4.} - \tilde{X}_{3.} = 2.74$	0.90	1.03	1.14
$\tilde{X}_{4.} - \tilde{X}_{2.} = 2.24$	0.85	0.85	1.14
$\tilde{X}_{2.} - \tilde{X}_{3.} = 0.50$	0.85	0.85	1.14

Table 7. Intermediate statistics of the one-stage procedure for the bacterial killing ability example

	Solvent 1	Solvent 2	Solvent 3	Solvent 4
N_i	24	36	67	16
S_i^2	2.10995	3.17085	5.88428	0.77969
U_i	0.04306	0.02778	0.01642	0.07700
V_i	0.03934	0.02778	0.01449	-0.15501
$\tilde{X}_{i.}$	97.4290	94.9047	94.8616	97.2339
		z* = 0.08808		

Table 8. The one-stage multiple comparison procedures for bacterial killing ability example

	test statistic		
pairwise difference	D1	NK1	T1
$\tilde{X}_{1.} - \tilde{X}_{3.} = 2.57$	0.97	1.20	1.20
$\tilde{X}_{1.} - \tilde{X}_{2.} = 2.53$	0.94	1.08	1.20
$\tilde{X}_{1.} - \tilde{X}_{4.} = 0.20$	0.90	0.90	1.20
$\tilde{X}_{4.} - \tilde{X}_{3.} = 2.37$	0.94	1.08	1.20
$\tilde{X}_{4.} - \tilde{X}_{2.} = 2.33$	0.90	0.90	1.20
$\tilde{X}_{2.} - \tilde{X}_{3.} = 0.04$	0.90	0.90	1.20

† D1 = One-stage Duncan type procedure.
NK1 = One-stage Newman-Keuls type procedure.
T1 = One-stage Tukey type procedure.

group and they are superior to solvent 2 and solvent 3. From this example we can clearly see that the one-stage procedure can both theoretically and practically provide a feasible solution to handle the difficulty of the two-stage procedure when an experiment is unexpectedly terminated earlier.

Acknowledgments

The authors are grateful to the reviewers for their valuable suggestions and for much helpful comments which improved the applicability and readability of the manuscript. The authors also would like to thank Ching-Feng Hsu and Feng-Yu Kao who, as part of their graduate studies, assisted in statistical computing.

References

[1] Carmer, S. G. and Swanson, M. R., "An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods," *Journal of the American Statistical Association*, Vol. 68, pp. 66–74 (1973).

[2] Bishop, T. A. and Dudewicz, E. J., "Exact Analysis of Variance with Unequal Variances: Test Procedures and Tables," *Technometrics*, Vol. 20, pp. 419–430 (1978).

[3] Chen, S. Y. and Chen, H. J., "A Range Test for the Equality of Means when Variances are Unequal," *American Journal of Mathematical and Management Sciences*, Vol. 20, pp. 145–170 (2000).

[4] Chen, S. Y., "One-Stage and Two-Stage Statistical Inference under Heteroscedasticity," *Communications in Statistics: Simulation and Computation*, Vol. 30, pp. 991–1009 (2001).

[5] Bishop, T. A., "Some Results on Simultaneous Inference for Analysis of Variance with Unequal Variances," *Technometrics*, Vol. 21, pp. 337–340 (1979).

[6] Hochberg, Y., "A Modification of the T-Method of Multiple Comparisons for a One-Way Layout with Unequal Variances," *Journal of the American Statistical Association*, Vol. 71, pp. 200–203 (1976).

[7] Tamhane, A. C., "Multiple Comparisons in Model I One-Way ANOVA with Unequal Variances," *Communications in Statistics - Theory and Method*, Vol. 6, pp. 15–32 (1977).

[8] Wilcox, R. R., "A Table of Percentage Points of the Range of Independent t Variables," *Technometrics*, Vol. 25, pp. 201–204 (1983).

[9] Chen, S. Y. and Chen, H. J. "Single-Stage Analysis of Variance under Heteroscedasticity," *Communications in Statistics - Simulation and Computation*, Vol. 27, pp. 641–666 (1998).

[10] Duncan, D. B., "Multiple Range and Multiple F Test," *Biometrics*, Vol. 1, pp. 1–42 (1955).

[11] Chen, H. J. and Lam, K. "Single-Stage Interval Estimation of the Largest Normal Mean under Heteroscedasticity," *Communications in Statistics - Theory and Methods*, Vol. 18, pp. 3703–3718 (1989).

[12] SAS Institute Inc., *SAS Language: Reference, Version 8*, Cary, NC. (1999).

[13] Chen, S. Y., Chen H. J. and Ding C. G., "An ANOVA Test for the Equivalency of Means under Unequal Variances," *Computational Statistics and Data Analysis*, Vol. 33, pp. 151–175 (2000).

Manuscript Received: Oct. 21, 2010
Accepted: Feb. 22, 2011