Testing Equality of Several Normal Means to a Specified Standard

K. KRISHNAMOORTHY

University of Southwestern Louisiana, Lafayette, LA 70504

ARVIND K. SHAH

University of South Alabama, Mobile, AL 36688

The problem of testing equality of several normal means to a specifed value when the population variances are unknown is considered. Two test procedures, one obtained by using Fisher's method of combining independent tests and the other based on the maximum of independent t statistics are proposed. In addition, the likelihood ratio test statistic and its asymptotic null distribution are derived. Power comparisons of these three test procedures are made using simulation. Assuming that the unknown variances are equal, these tests are again compared with an appropriately modified analysis of variance procedure. Based on the simulation study, the likehood ratio test and Fisher's method are recommended for use in general. When the variances are equal, the modified analysis of variance test is superior.

Introduction

Do any of several population means differ from a specified standard or target value? Mee, Shah, and Lefante (1987) addressed this question and presented an inference procedure for the case of normally distributed populations with unknown but equal variances. Since the assumption of equal variances is not always practical (for example, due to variation in manufacturing equipments, locations, or operating environments), this paper presents inference procedures for the case of normally distributed populations with unknown and unequal variances. It is of interest to note here that if the standard and the variances are unknown, then the present problem reduces to the well-known Behrens-Fisher problem. (See Taneja and Dudewicz (1993, pp. 447 - 477), Lehmann (1986, p. 304), Pfanzagl (1974), and Asiribo and Garland (1989)). For multiple comparisons with the control population, refer to Dudewicz and Dalal (1983) and Bofinger and Lewis (1992, pp. 25 - 45). Also, if the group means are known to be equal and we want to test or estimate the common mean then the problem is known in the literature as the "common mean problem", and we refer the readers to Cohen and Sackrowitz (1977, 1984).

The present situation arises in many natural settings. For example, a drug manufacturing company may wish to test whether different machines on production lines dispense a particular ingredient in the same average pre-specified quantity. Maxcy and Lowry (1984) presented several data sets from filling operations of ground beef packages, milk cartons and carbonated beverages. They used analysis of variance (AOV) and tested for equality of group means instead of comparing group means to a set-forth standard. Mee, Shah, and Lefante (1987) presented an illustrative example which consists of four machines producing ball bearings with an intended diameter of one millimeter. The authors used the target value in computing the numerator sum of squares of the F test statistic in the AOV procedure and tested the equality of the means to the target value. This is an appropriately modified analysis of variance procedure for the present problem; we refer to it as MAOV. However, application of Hartley's $F_{\max}$ test to Maxey and Lowry data sets indicates significant differences among group variances, and hence, AOV and its modified version MAOV (as they assume equal group variances) are not appropriate testing methods for their data sets.

Dr. Krishnamoorthy is an Assistant Professor in the Department of Statistics.

Dr. Shah is Professor of Statistics in the Department of Mathematics and Statistics.

In the following section, we propose two procedures for testing the equality of the means to the specified value when the variances are unknown. The first one is based on the maximum of independent t test statistics (MAXT). The second one is obtained by combining independent t tests using Fisher's method (FMCI). Then we derive the likelihood ratio test (LRT) statistic and its moments. Although the exact null density of LRT statistic can be obtained using its moments, it can be quite complicated making the computation of p values very difficult. Hence, we give its asymptotic null distribution so that the p values can be computed easily and with sufficient accuracy.

It is to be noted that all of these test procedures may be asymptotically equally efficient. In order to understand the small sample properties of these tests, their powers are computed through simulation. We observe from the power comparison that the LRT and the FMCI seem to be superior in general when the variances are unequal. When the variances are equal the MAOV method performs better. Lastly, these test procedures are illustrated using two examples.

Combined Test Procedures

Let $x_{i1}, x_{i2}, \ldots, x_{in_i}$ be independent observations from the i^{th} normal population with mean μ_i and variance σ_i^2, $i = 1, 2, \ldots, k$. Assume that all the means and the variances are unknown. We want to test

$$\begin{aligned} H_0&: \quad \mu_1 = \mu_2 = \cdots = \mu_k = \mu_0 \\ &\text{versus} \\ H_a&: \quad \mu_i \neq \mu_0 \quad \text{for some } i \end{aligned} \tag{1}$$

where μ_0 is a known standard value.

Let $\bar{x}_i$ and $s_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2/(n_i - 1)$ denote, respectively, the i^{th} sample mean and variance. Then $T_i = \sqrt{n_i}(\bar{x}_i - \mu_0)/s_i$ is the t test statistic for testing $\mu_i = \mu_0$. Note that T_i's are independent having the t distribution with $n_i - 1$ degrees of freedom.

Test Based on Independent t Tests (MAXT)

Let c_i denote the $100\{[1 + (1-\alpha)^{1/k}]/2\}^{\text{th}}$ percentile point of $T_i, i = 1, 2, \ldots, k$. As T_i's are independent,

$$\Pr(|T_1| \leq c_1, |T_2| \leq c_2, \ldots, |T_k| \leq c_k) = \prod_{i=1}^{k} \Pr(|T_i| \leq c_i) = 1 - \alpha. \tag{2}$$

Hence, the intersection of the intervals $\bar{x}_i \pm c_i s_i/\sqrt{n_i}$ is an exact $100(1-\alpha)\%$ confidence interval for the unknown common mean of all the populations. Therefore, the test that rejects H_0 if μ_0 does not belong to

$$\left[\max_{1\leq i\leq k} (\bar{x}_i - c_i s_i/\sqrt{n_i}),\ \min_{1\leq i\leq k} (\bar{x}_i + c_i s_i/\sqrt{n_i})\right] \tag{3}$$

or if the interval is empty, has exact size α. If all the sample sizes are equal, then $c_i = c$ for all i, and the test in (3) will be equivalent to rejecting H_0, if

$$T_{(k)} = \max_{1\leq i\leq k} |T_i| > c.$$

It follows from (2) that this test procedure not only tests H_0 but also points out which population means are significantly different from the standard. For example, if $\bar{x}_i \pm c_i s_i/\sqrt{n_i}$ does not contain μ_0, then we conclude that the i^{th} population mean is significantly different from the standard μ_0. The values of c_i can be computed using commercially available statistical software, electronic tables, or special calculators giving percentiles for the t distribution. These include SAS, StaTable, and Hewlett Packard calculator 21S, among many others.

Next, we combine the p values of the independent t test statistics using Fisher's (1950, p. 99) method to obtain a chi squared statistic for testing the hypotheses in (1).

Test Procedure FMCI

Let p_i denote the p value of the Student's t test based on the i^{th} sample alone for testing

$$\begin{aligned} H_0&: \quad \mu_i = \mu_0 \\ &\text{versus} \\ H_a&: \quad \mu_i \neq \mu_0. \end{aligned} \tag{4}$$

It can be easily verified, using the probability integral transform, that under H_0 in (1), $-2\,\ell\text{n}p_1$, $-2\,\ell\text{n}p_2$, $\ldots$, $-2\,\ell\text{n}p_k$ are independent having the chi squared distribution with two degrees of freedom. Hence,

$$V = -2 \sum_{i=1}^{k} \ell\text{n}p_i \tag{5}$$

follows the chi squared distribution with $2k$ degrees of freedom. Thus, the decision rule is to reject H_0 in (1) if

$$V \geq \chi^2_{2k}(\alpha)$$

where $\chi^2_{2k}(\alpha)$ denotes the $100(1-\alpha)^{\text{th}}$ percentile point of the chi squared distribution with $2k$ degrees of freedom.

Likelihood Ratio Test (LRT)

Let $a_i^2 = (n_i - 1)s_i^2$, $i = 1, 2, \ldots, k$. Noticing that the maximum likelihood estimator of σ_i^2 under H_0 in (1) is $[a_i^2 + n_i(\bar{x}_i - \mu_0)^2]/n_i$, it can be easily verified that the likelihood ratio test statistic for testing (1) is

$$\Lambda = \prod_{i=1}^{k} \left[a_i^2/(a_i^2 + n_i(\bar{x}_i - \mu_0)^2)\right]^{n_i/2} = \prod_{i=1}^{k} (V_i)^{n_i/2} \text{ (say).} \tag{6}$$

In order to obtain a decision rule regarding the hypotheses in (1), we need the probability distribution of Λ or any one-to-one function of Λ. As noted earlier, it is difficult to find the exact distribution of Λ, and hence, we give the well-known Box (1949) series approximation, discussed in Anderson (1984, p. 311), to the distribution of Λ. To use Box's method, all the moments of Λ are needed. These moments can be computed using the fact that $V_1, V_2, \ldots, V_k$ in (6) are independent and under H_0, the V_i follow the beta distribution $\beta((n_i-1)/2, 1/2)$. Indeed, after evaluating $E(\Lambda^h)$ for $h = 1, 2, \ldots$, and then proceeding along the lines of Anderson (1984, p. 311), it can be shown that

$$\begin{aligned}\Pr(-2\rho\,\ell\text{n}\Lambda \leq u) = {} & \\ \Pr(\chi_k^2 \leq u) + {} & \{w_2[\Pr(\chi^2_{k+4} \leq u) - \Pr(\chi_k^2 \leq u)]\} \\ & + \{w_3[\Pr(\chi^2_{k+6} \leq u) - \Pr(\chi_k^2 \leq u)]\} + R\end{aligned} \tag{7}$$

where χ_b^2 denotes the chi squared random variable with b degrees of freedom, R is the remainder term of the order $(\sum_i 1/n_i)^4$, and, for $i = 1, 2, \ldots, k$,

$$\begin{aligned}\rho &= (1 - 3n_0/(2k)) \\ n_0 &= \sum\nolimits_i (1/n_i) \\ w_2 &= \left(-\frac{9n_0^2}{16k} + \frac{1}{2}\sum\nolimits_i \frac{1}{n_i^2}\right)/\rho^2 \\ w_3 &= \left(\frac{9n_0^3}{8k^2} - \frac{3n_0}{2k}\sum\nolimits_i \frac{1}{n_i^2} + \frac{3}{8}\sum\nolimits_i \frac{1}{n_i^3}\right)/\rho^3.\end{aligned} \tag{8}$$

If $n_1 = n_2 = \cdots = n_k = n$, then $\rho = (1 - 3/2n)$, $w_3 = 0$, and $w_2 = -k/[4(2n-3)^2]$. In this case,

$$\begin{aligned}\Pr(-2\rho\,\ell\text{n}\Lambda \leq u) = {} & \\ \Pr(\chi_k^2 \leq u) + {} & \{w_2[\Pr(\chi^2_{k+4} \leq u) - \Pr(\chi_k^2 \leq u)]\} + R\end{aligned} \tag{9}$$

and R is the error term of the order n^{-4}.

Note, that for given k and sample sizes, a computer is needed to compute the percentile points using (7) or (9). However the p values, which are more important than the percentile points, can be computed using electronic calculators or available tables for the chi squared distribution.

Also note that the percentile points obtained through the Box series approximation are good enough for practical use, and more accurate points can be obtained using more terms in the series expansion. We checked the accuracy of the percentile points obtained through the Box series approximation with those computed from the exact probability distribution. These exact and approximated percentile points generally agree very well to four decimal places for sample sizes above ten and agree to two decimal places for very small samples. For example, when $(n_1, n_2) = (5, 7)$ the approximated and exact 95^{th} percentile points are 5.9150413 and 5.9150823; and when $(n_1, n_2) = (4, 4)$, then the approximated and exact 95^{th} percentile points are 5.7524984 and 5.7503243, respectively.

We note that the two-stage multiple comparisons procedure with the control population given by Dudewicz and Dalal (1983) can be modified, by taking the initial sample size from the control population as infinity, to handle the present problem. However, for the present problem, the methods proposed in this paper are simpler to use than the modified two-stage procedure, and hence, it is not discussed here.

Power Simulation and Conclusion

The powers of the test procedures given in the preceding sections are estimated, at the level of significance 0.05, using 100,000 simulated samples of different sizes when $k = 3$ and 4. The RNNOA subroutine of IMSL (1989) was used to generate samples from the normal distribution. This subroutine is based on an acceptance/rejection technique due to Kinderman and Ramage (1976). The same 100,000 samples were used to estimate the power of all four tests at a given set of parameter points. As these are the estimates of probabilities, the estimated values of the standard errors are less than or equal to $\sqrt{0.25/100,000} = 0.0016$. Hence, these estimated powers are accurate up to at least two decimal places. To estimate the powers, without loss of generality, we assumed that the specified value of the common mean $\mu_0 = 0$. Tables 1, 2 and 5 present the powers for a case of unequal variances. It appears from these tables that only a minor difference exists between the powers of the LRT and the FMCI. When only one group mean is different from zero, the test MAXT outperforms the other two tests and the LRT outperforms the FMCI as the magnitude of the difference increases. On the other hand, when more than one group mean is different from zero, MAXT has lower power than both LRT and FMCI. Based on the power comparisons and the trade-offs in power under various configurations, the LRT and FMCI seem to be more appropriate for general usage.

When $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, Mee, Shah and Lefante (1987) proposed the test statistic

$$F_0 = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \mu_0)^2/k}{\sum_{i=1}^{k} (n_i - 1)s_i^2/(N-k)} \tag{10}$$

Table 1. Simulated Powers of the Tests at $\alpha = 0.05$, $\mu_0 = 0$, $k = 3$, $n_1 = 18$, $n_2 = 12$, $n_3 = 16$, $\sigma_1^2 = 2$, $\sigma_2^2 = 4$, and $\sigma_3^2 = 7$

(μ_1, μ_2, μ_3)	FMCI	LRT	MAXT
(0.0, 0.0, 0.0)	0.05	0.05	0.05
(0.0, 0.0, 0.5)	0.08	0.08	0.74
(0.0, 0.0, 1.0)	0.17	0.18	0.18
(0.0, 0.0, 2.0)	0.59	0.62	0.65
(0.0, 0.0, 3.0)	0.92	0.93	0.96
(0.0, 0.0, 4.0)	0.99	1.00	1.00
(0.0, 0.5, 0.5)	0.13	0.13	0.12
(0.0, 0.5, 1.0)	0.24	0.24	0.21
(0.0, 1.0, 1.0)	0.39	0.40	0.33
(0.0, 1.0, 1.5)	0.58	0.58	0.50
(0.0, 1.5, 1.5)	0.77	0.77	0.67
(0.0, 2.0, 2.5)	0.99	0.99	0.96
(0.5, 0.5, 0.5)	0.29	0.29	0.24
(0.5, 0.5, 1.0)	0.42	0.41	0.33
(0.5, 1.0, 1.0)	0.57	0.56	0.43
(0.5, 1.0, 2.0)	0.86	0.85	0.75
(1.0, 1.0, 2.0)	0.97	0.97	0.90
critical values	12.60	7.81	

Table 2. Simulated Powers of the Tests at $\alpha = 0.05$, $\mu_0 = 0, k = 4$, $n_1 = 10$, $n_2 = 13$, $n_3 = 20$, $n_4 = 15$, $\sigma_1^2 = 2$, $\sigma_2^2 = 4$, $\sigma_3^2 = 7$, and $\sigma_4^2 = 12$

$(\mu_1, \mu_2, \mu_3, \mu_4)$	FMCI	LRT	MAXT
(0, 0, 0, 0)	0.05	0.05	0.05
(0, 0, 0, 1)	0.11	0.11	0.11
(0, 0, 0, 2)	0.30	0.31	0.34
(0, 0, 0, 3)	0.61	0.64	0.70
(0, 0, 0, 4)	0.85	0.88	0.93
(0, 0, 0, 5)	0.97	0.98	0.99
(0, 0, 1, 1)	0.28	0.28	0.25
(0, 0, 1, 2)	0.52	0.52	0.45
(0, 0, 2, 2)	0.87	0.87	0.82
(0, 0, 2, 3)	0.96	0.96	0.92
(0, 1, 1, 1)	0.51	0.50	0.38
(0, 1, 1, 2)	0.71	0.70	0.54
(0, 1, 2, 2)	0.94	0.93	0.85
(1, 1, 1, 1)	0.75	0.74	0.54
(1, 2, 1, 2)	0.98	0.98	0.89
(1, 2, 1, 3)	1.00	1.00	0.95
critical values	15.51	9.48	

Table 3. Simulated Powers of the Tests at $\alpha = 0.05$, $\mu_0 = 0$, $k = 3$, $n_1 = 10$, $n_2 = 13$, $n_3 = 20$, and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 5$

(μ_1, μ_2, μ_3)	FMCI	LRT	MAXT	MAOV
(0.0, 0.0, 0.0)	0.05	0.05	0.05	0.05
(0.0, 0.0, 0.5)	0.10	0.10	0.10	0.11
(0.0, 0.0, 1.0)	0.30	0.30	0.32	0.33
(0.0, 0.0, 2.0)	0.85	0.86	0.91	0.91
(0.0, 0.0, 3.0)	1.00	1.00	1.00	1.00
(0.0, 0.5, 0.5)	0.15	0.14	0.13	0.15
(0.0, 0.5, 1.0)	0.36	0.34	0.34	0.38
(0.0, 1.0, 1.0)	0.50	0.49	0.42	0.52
(0.0, 1.0, 1.5)	0.76	0.75	0.72	0.79
(0.0, 1.5, 1.5)	0.87	0.87	0.80	0.89
(0.0, 2.0, 2.5)	1.00	1.00	1.00	1.00
(0.1, 0.5, 0.5)	0.15	0.14	0.13	0.15
(0.8, 0.9, 1.0)	0.57	0.56	0.44	0.57
(0.5, 1.0, 1.5)	0.78	0.77	0.72	0.80
(1.0, 1.5, 2.0)	0.98	0.98	0.95	0.98
critical values	12.60	7.81		2.84

Table 4. Simulated Powers of the Tests at $\alpha = 0.05$, $\mu_0 = 0$, $k = 4$, $n_1 = 8$, $n_2 = 14$, $n_3 = 11$, $n_4 = 20$, and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 7$

$(\mu_1, \mu_2, \mu_3, \mu_4)$	FMCI	LRT	MAXT	MAOV
(0, 0, 0, 0)	0.05	0.05	0.05	0.05
(1, 0, 0, 0)	0.10	0.10	0.09	0.11
(2, 0, 0, 0)	0.22	0.25	0.24	0.33
(3, 0, 0, 0)	0.44	0.51	0.53	0.69
(4, 0, 0, 0)	0.66	0.76	0.81	0.92
(1, 1, 0, 0)	0.20	0.21	0.18	0.24
(1, 2, 0, 0)	0.53	0.54	0.54	0.62
(1, 0, 3, 0)	0.72	0.75	0.77	0.86
(4, 0, 0, 3)	1.00	1.00	1.00	1.00
(0, 1, 2, 1)	0.68	0.67	0.53	0.72
(0, 1, 4, 1)	0.97	0.97	0.97	0.99
(0, 1, 4, 4)	1.00	1.00	1.00	1.00
(1, 1, 1, 1)	0.52	0.51	0.35	0.53
(1, 2, 2, 1)	0.91	0.90	0.76	0.93
(1, 2, 2, 2)	0.99	0.98	0.92	0.99
(5, 1, 1, 1)	0.98	0.98	0.96	1.00
critical values	15.51	9.49		2.56

where $N = \sum_{i=1}^{k} n_i$, for testing the hypotheses (1). Note that, when H_0 is true, F_0 follows the F distribution with k and $(N-k)$ degrees of freedom. We have referenced this test procedure as MAOV earlier. The powers of the test procedures LRT, MAXT, FMCI, and MAOV are estimated through simulation and are given in Tables 3 and 4. As expected, when $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, the test MAOV is superior. Even in this case, LRT and FMCI (which are developed without using the knowledge of equal variances) perform quite well.

It should be recalled here that all the tests except LRT are exact. Even LRT is almost exact since the critical values are almost exact. In Tables 1 through 4, we have included the configuration of (μ_1, μ_2, μ_3) = (0,0,0) in order to show that the estimated level of these tests under the null hypothesis of $\mu_1 = \mu_2 = \mu_3 = 0$ is almost equal to the prespecified nominal level. As expected, the estimated levels are equal to the set nominal level of 0.05, which also indicates the validity of simulation.

Illustrative Examples

The test procedures developed here will now be illustrated on the numerical data sets reported by Romano (1977, p. 248) and Maxcy and Lowry (1984).

We first consider Romano's example. Each observation is multiplied by 1,000 for simplicity. In an industrial corporation, four production lines are set to produce ball bearings with a diameter of 1 mm. At the end of a day's production, ten ball bearings are randomly and independently selected from each of the four lots produced by these production lines. The ball bearing diameters, in mm, are: line 1: 1.18, 1.42, 0.69, 0.88, 1.62, 1.09, 1.53, 1.02, 1.19, 1.32; line 2: 1.72, 1.62, 1.69, 0.79, 1.79, 0.77, 1.44, 1.29, 1.96, 0.99; line 3: 0.58, 1.37, 0.83, 1.38, 1.62, 1.16, 1.78, 1.14, 0.64, 0.79; and line 4: 1.01, 1.46, 1.21, 1.16, 1.48, 0.67, 1.03, 1.13, 1.40, 1.21. From these data, $\bar{x}_1 = 1.194$, $\bar{x}_2 = 1.406$, $\bar{x}_3 = 1.129$, $\bar{x}_4 = 1.176$, $s_1^2 = 0.083916$, $s_2^2 = 0.183449$, $s_3^2 = 0.170210$, $s_4^2 = 0.059204$. The normality assumption was tested using the Shapiro-Wilk statistic which gave p values of 0.9835, 0.2834, 0.6586, and 0.4291, for samples 1 through 4, respectively. Hartley's $F_{\max}$ test does not provide sufficient evidence against the assumption of homogeneity of variance at 5% level of significance.

Table 5. Simulated Powers of the Tests at $\alpha = 0.05$, $\mu_0 = 0$, $k = 3$, $n_1 = n_2 = n_3 = 10$, $\sigma_1^2 = 2$, $\sigma_2^2 = 4$, and $\sigma_3^2 = 7$

(μ_1, μ_2, μ_3)	FMCI	LRT	MAXT
(0.0, 0.0, 0.0)	0.05	0.05	0.05
(0.0, 0.0, 0.5)	0.07	0.07	0.06
(0.0, 0.0, 1.0)	0.12	0.12	0.12
(0.0, 0.0, 2.0)	0.35	0.36	0.38
(0.0, 0.0, 3.0)	0.66	0.69	0.74
(0.0, 0.0. 5.0)	0.98	0.99	0.99
(0.0, 0.5, 0.5)	0.10	0.10	0.09
(0.0, 0.5, 1.0)	0.16	0.16	0.14
(0.0, 1.0, 1.0)	0.28	0.28	0.24
(0.0, 1.0, 1.5)	0.41	0.40	0.33
(0.0, 1.5, 1.5)	0.58	0.58	0.48
(0.0, 2.0, 1.5)	0.76	0.76	0.69
(0.0, 1.5, 2.0)	0.71	0.71	0.59
(0.0, 2.0, 2.0)	0.85	0.85	0.75
(0.0, 2.5, 2.5)	0.97	0.97	0.92
(0.0, 3.0, 2.0)	0.98	0.98	0.96
(0.0, 3.0, 3.0)	1.00	1.00	0.98
(0.0, 4.0, 3.0)	1.00	1.00	1.00
(0.5, 0.5, 0.5)	0.18	0.17	0.14
(1.0, 0.5, 1.0)	0.48	0.48	0.40
(1.0, 2.0, 1.0)	0.88	0.87	0.75
(2.0, 2.0, 3.0)	1.00	1.00	0.99
critical values	12.60	7.78	2.92

Romano applied the usual AOV procedure and concluded that there are no significant differences among the lot means. Mee, Shah, and Lefante (1987) considered the same example and computed the statistic

$$F_0 = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \mu_0)^2/k}{\sum_{i-1}^{k} a_i^2/(N-k)} = 5.03.$$

Noting that F_0 follows the F distribution with $k = 4$ and $N - k = 36$ degrees of freedom, $F_0 = 5.03$ corresponds to the p value of 0.0025, indicating that at least one process is out of control and with a mean other than 1 mm. The multiple comparison procedure outlined in Mee, Shah, and Lefante declared the second machine to be out of control. Because of the pooling of variances across samples, the MAOV procedure is expected to have higher power (and, hence, lower p values) than the procedures discussed in this paper, which are specifically developed for the unequal variance case. This, indeed, is the case and is demonstrated next. For Romano's example, the LRT statistic $-2\rho\,\ell\text{n}\Lambda$ reduces to 14.09, which corresponds to the p value of 0.0068. Here, $|T_1| = 2.1178$, $|T_2| = 2.9976$, $|T_3| = 0.9888$, $|T_4| = 2.2874$, $p_1 = 0.0633$, $p_2 = 0.0150$, $p_3 = 0.3486$, $p_4 = 0.0480$. Under the MAXT procedure, the p value $= \Pr(T_{(4)} > 2.9976) = 1 - \prod_{i=1}^{4}[2F_i(2.9976) - 1] = 0.0587$. (Here $F_i(x)$ denotes the cumulative probability under the t distribution with $n_i - 1 = 9$ degrees of freedom.) Under the FMCI procedure, $V = -2\,\ell\text{n}(p_1p_2p_3p_4) = 22.10$ and p value $= \Pr(\chi_8^2 \geq 22.10) = 0.0047$.

We next illustrate this test procedure on data from the filling operation presented in Maxcy and Lowry (1984). This data set represents the gross weight (in grams) of samples consisting of 10 packages of ground beef. For brevity, we take group numbers of 5, 9 and 11 from their Table 2. The data are: group 5: 1410.0, 1393.9, 1405.9, 1404.2, 1387.3, 1398.5, 1399.9, 1392.5, 1402.5, 1391.8; group 9: 1388.2, 1382.9, 1395.6, 1388.2, 1389.7, 1391.4, 1383.3, 1390.5, 1398.4, 1390.6; and group 11: 1409.9, 1386.7, 1410.5, 1401.0, 1413.0, 1387.6, 1412.0, 1390.7, 1417.2, and 1383.9. These data yield the means 1398.65, 1389.88, and 1401.25 gm and standard deviations 7.19, 4.79, and 12.82 gm, respectively. The normality assumption was tested using the Shapiro-Wilk statistic, which produced p values of 0.9400, 0.5823, and 0.0970 for groups 5, 9, and 11, respectively. Hartley's $F_{\max}$ test indicates that there is a significant difference among the variances associated with the three filling operations at the 5% level. Assuming that these filling machines are set to fill 1400 gm per package, the LRT statistic $-2\rho\,\ell\text{n}\Lambda$ simplifies to 15.5990 (p value $= 0.0013$). Here $|T_1| = 0.59$, $|T_2| = 6.68$, $|T_3| = 0.31$, $p_1 = 0.5697$, $p_2 = 9.06\times10^{-5}$, and $p_3 = 0.7636$. Under the MAXT procedure, the p value $= \Pr(T_{(3)} > 6.68) = 0.00027$. Under the FMCI procedure, $V = 20.28$ and the p value $= \Pr(\chi_6^2 \geq 20.28) = 0.0025$. Each of the three test procedures provides strong evidence to conclude that at least one of the machines is filling packages at a mean weight of other than 1400 gm. Using the MAXT procedure with $\alpha = 0.05$ and $c = 2.92287$, the 95% simultaneous confidence interval for μ_5, μ_9, μ_{11} is [(1392.00, 1405.30), (1385.45, 1394.3), (1389.40, 1413.10)]. From this interval, we can conclude that the ninth machine is out of control (i.e., $\mu_9 \neq 1400$ gm).

Acknowledgments

We are grateful to the editor and two anonymous referees for their helpful comments and suggestions which improved the final version of this paper. The work of the first author was supported in part by a Faculty Summer Research Grant (1993) of the University of Southwestern Louisiana.

References

ANDERSON, T.W. (1984). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York, NY.

ASIRIBO, O. and GARLAND, J. (1989). "Some Simple Approximate Solutions to the Behrens-Fisher Problem". *Communications in Statistics -Theory and Methods* 18, pp. 1201-1216.

BOFINGER, E. and LEWIS, G. J. (1992). *The Frontiers of Modern Statistical Inference Procedures,* Vol. II edited by E. Bofinger, E. J. Dudewicz, G. J. Lewis, and K. Mengersen. American Sciences Press, Columbus, OH.

BOX, G. E. P. (1949). "A General Distribution Theory for a Class of Likelihood Criteria". *Biometrika* 36, pp. 317-346.

COHEN, A. and SACKROWITZ, H. B. (1977). "Hypothesis Testing for the Common Mean and for Balanced Incomplete Block Designs". *The Annals of Statistics* 5, pp. 1195-1211.

COHEN, A. and SACKROWITZ, H. B. (1984). "Testing Hypothesis about Common Mean of Two Normal Populations". *Journal of Statistical Planning and Inference* 9, pp. 207-227.

DUDEWICZ, E. J. and DALAL, S. R. (1983). "Multiple Comparisons with a Control when Variances are Unknown and Unequal". *American Journal of Mathematical and Management Sciences* 3, pp. 275-295.

FISHER, R. A. (1950). *Statistical Methods for Research Workers,* 11th ed. Oliver & Boyd, London, England.

IMSL (1989). *User's Manual - FORTRAN Subroutines for Statistical Analysis* Vol. 3, International Mathematical Statistical Library, Houston, TX.

KINDERMAN, A. J. and RAMAGE, J. G. (1976). "Computer Generation of Normal Random Variables". *Journal of the American Statistical Association* 71, pp. 893-896.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses.* John Wiley & Sons, New York, NY.

MAXCY, R. B. and LOWRY, S. R. (1984). "Evaluating Variability of Filling Operations". *Food Technology* 38, pp. 51-55.

MEE, R. W.; SHAH, A. K.; and LEFANTE, J. J. (1987). "Comparing *k* Independent Samples with Known Standard". *Journal of Quality Technology* 19, pp. 75-81.

PFANZAGL, J. (1974). "On the Behrens-Fisher Problem". *Biometrika* 61, pp. 39-47.

ROMANO, A. (1977). *Applied Statistics for Science and Industry.* Allyn and Bacon, Boston, MA.

TANEJA, B. K. and DUDEWICZ, E. J. (1993). *Multiple Comparisons, Selection, and Applications in Biometry* edited by F. M. Hoppe. Marcel Dekker, New York, NY.

Key Words: *Likelihood Ratio Tests, Maximum Likelihood, Moment Estimation, Tests for Normality.*