Determining Sample Sizes for Precise Contrast Analysis With Heterogeneous Variances

Show-Li Jan and Gwowen Shieh

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2014 39: 91
DOI: 10.3102/1076998614523069

Journal of Educational and Behavioral Statistics
2014, Vol. 39, No. 2, pp. 91–116
DOI: 10.3102/1076998614523069

Determining Sample Sizes for Precise Contrast Analysis With Heterogeneous Variances

Show-Li Jan
Chung Yuan Christian University

Gwowen Shieh
National Chiao Tung University

The analysis of variance (ANOVA) is one of the most frequently used statistical analyses in practical applications. Accordingly, the single and multiple comparison procedures are frequently applied to assess the differences among mean effects. However, the underlying assumption of homogeneous variances may not always be tenable. This study examines the sample size procedures for precise interval estimation of linear contrasts within the context of one-way heteroscedastic ANOVA models. The desired precision of both individual and simultaneous confidence intervals is evaluated with respect to the control of expected half width and to the tolerance probability of interval half width within a designated value. Supplementary computer programs are developed to aid the usefulness and implementation of the proposed techniques. The suggested sample size procedures improve upon the existing approaches and extend the methodology development in the statistical literature.

Keywords: *confidence interval, contrast, precision, sample size*

Individual and multiple comparisons of mean effects in homoscedastic analysis of variance (ANOVA) models have received considerable attention in the literature. Accordingly, Bird (2004); Bretz, Hothorn, and Westfall (2010); Cumming (2012); Hahn and Meeker (1991); Hochberg and Tamhane (1987); Hsu (1996); Smithson (2003); Westfall, Tobias, Rom, Wolfinger, and Hochberg (2011); and the references therein provide an excellent and thorough account of the associated properties and explications for constructing confidence intervals in ANOVA and related models. Although the homogeneity of variance formulation provides a convenient and useful setup, it is not unusual for the homoscedasticity assumption to be violated in actual applications. Specifically, Fenstad (1983), Grissom (2000), and Wilcox (1987) emphasized that there are theoretical reasons to expect and empirical results to document the existence of heteroscedasticity is more common than most researchers realize. Therefore, it is prudent to

recommend employing suitable techniques that are superior to the traditional inferential methods under various conditions of heteroscedasticity.

The comparisons of mean effects require the formulation of a contrast, which is a linear combinations of population means and the coefficients of the means add up to zero. The difference between two group means or pairwise comparison is the simplest case of a linear contrast, whereas a complex comparison may involve several treatment means and designated coefficients in order to address theoretically and practically meaningful questions. Under the independence, normality, and homogeneity of variance assumptions in ANOVA, the inference for a linear contrast of mean effects can be conducted with a single degree of freedom F statistic or a t statistic. However, it is often desirable and sensible to perform multiple comparisons among means through a family of confidence intervals for contrasts to provide specific answers to critical research questions. Thus, it becomes necessary to consider simultaneous interval procedures that permit the family confidence coefficient to be controlled. Specifically, the Bonferroni procedure is useful when the number of comparisons to be investigated is identified in advance of the study. Whereas the Scheffe (1959), Tukey (1994), and Kramer (1956) methods are applicable for multiple comparisons of planned and post hoc contrasts and still maintain the desired overall confidence level of the joint confidence intervals. Furthermore, comprehensive guidelines and practical implications can be found in Kutner, Nachtsheim, Neter, and Li (2005) and Maxwell and Delaney (2004).

In order to take into account the heteroscedastic situations, particular emphasis is devoted to the problem of mean comparisons when the population variances are unknown and cannot be assumed equal. As shown in Rossi (1975), the approximate t-solution of Welch (1947) can be readily applied to construct confidence intervals of linear contrasts involved more than two mean effects. The intrinsic notion is a generalization of the approach suggested independently by Satterthwaite (1946), Smith (1936), and Welch (1938) for the Behrens–Fisher problem of comparing the means of two populations. For pairwise and general multiple comparisons of means with unequal variances, Dunnett (1980) and Tamhane (1979) described several feasible procedures and compared their performance of confidence levels and interval widths by Monte Carlo simulation study. In view of the overall behavior and computational requirement, the six methods considered in Brown and Forsythe (1974), Dunnett (1980), Games and Howell (1976), Tamhane (1977), and Ury and Wiggins (1971) are potentially appropriate for practical applications.

The reporting of effect sizes and associated confidence intervals for primary results in all empirical social science research has been recommended in Wilkinson and American Psychological Association Task Force on Statistical Inference (1999), the American Educational Research Association Task Force on Reporting of Research Methods (2006), and the *Publication Manual of the American Psychological Association* (American Psychological Association,

2010). According to the editorial guidelines and methodological recommendations of several prominent educational and psychological journals, it is necessary to include some measures of effect size and confidence intervals in all research studies (Alhija & Levy, 2009; Cohen, 1990, 1994; Dunst & Hamby, 2012; Fritz, Morris, & Richler, 2012; Odgaard & Fowler, 2010; Sun, Pan, & Wang, 2010). Within the context of ANOVA, the use of effect sizes in conjunction with confidence intervals has been emphasized in Bird (2002); Levine, Weber, Park, and Hullett (2008); and Robey (2004). It is essential to note that a linear contrast between two or more means can be considered an effect size index in the individual and multiple comparison investigations. For advance research design planning, the methods for computing necessary sample sizes of desired confidence intervals of linear contrasts for multiple comparison studies have been presented in Pan and Kupper (1999). However, it is important to note that their methods are confined to the homogeneous variance and balanced design. In view of the continued recommendation for the use of confidence intervals in all empirical studies, this study aims to expedite this research practice by presenting the sample size procedures for precise individual and simultaneous confidence intervals for single and multiple comparisons in fixed-effects heteroscedastic ANOVA designs.

Specifically, the individual comparison of contrasts and six renowned multiple comparison methods will be considered under the general framework of heterogeneous variances and unbalanced structures. In addition, the desired precision of a confidence interval is assessed with respect to the control of expected half width and to the tolerance probability of interval half width within a designated value. Hence, the proposed sample size calculations for precise interval estimation are described in terms of two distinct features. One method gives the minimum sample size, such that the expected half widths of a family of confidence intervals are within the designated bounds. The other provides the sample size needed to guarantee, with a given tolerance probability, that the half widths of a family of confidence intervals will not exceed the planned ranges. The notion of expected half width for sample size calculations is frequently introduced in standard texts. However, considerable attention has focused on the criterion of tolerance probability of interval half width within a given value. For example, see Kelley, Maxwell, and Rausch (2003); Kupper and Hefner (1989); and Liu (2009) for related discussion in the context of estimating the mean difference between two normal populations with homoscedasticity. Consequently, this investigation updates and expands the current work in sample size determinations of confidence interval estimation for mean comparisons in ANOVA, especially the existing results in Kupper and Hafner (1989), Pan and Kupper (1999), Shieh and Jan (2012), and Wang and Kupper (1997). In addition, the computations of these procedures involve iterative algorithms not currently available in statistical packages, and hence, the computer codes are presented to facilitate the recommended approaches for computing the necessary sample sizes of linear contrast confidence intervals with designated precision in planning research designs.

Individual Comparisons of Means

Consider the one-way heteroscedastic ANOVA model in which the observations Y_{ij} are assumed to be independent and normally distributed with expected values μ_i and variances σ_i^2:

$$Y_{ij} \sim N(\mu_i, \sigma_i^2), \tag{1}$$

where μ_i and σ_i^2 are unknown parameters, $i = 1, \ldots, g\ (\geq 2)$ and $j = 1, \ldots, N_i$. For inference purposes of linear combinations of mean parameters, a contrast is defined as

$$\psi = \sum_{i=1}^{g} c_i \mu_i,$$

where c_i are the contrast coefficients with $\sum_{i=1}^{g} c_i = 0$. It follows from the model assumption in Equation 1 that a convenient unbiased contrast estimator $\widehat{\psi}$ for ψ is of the form

$$\widehat{\psi} = \sum_{i=1}^{g} c_i \bar{Y}_i,$$

where $\bar{Y} = \sum_{j=1}^{N_i} Y_{ij}/N_i$ is the ith group sample mean and is an unbiased estimator of μ_i for $i = 1, \ldots, g$. Moreover, the linear estimator $\widehat{\psi}$ has the distribution

$$\widehat{\psi} \sim N(\psi, \Sigma), \tag{2}$$

where $\Sigma = \mathrm{Var}(\widehat{\psi}) = \sum_{i=1}^{g} c^2 \sigma_i^2 / N_i$. Also, an unbiased estimator $\widehat{\Sigma}$ of Σ can be obtained by replacing the variance σ_i^2 in Σ with its unbiased estimator S_i^2 as follows:

$$\widehat{\Sigma} = \sum_{i=1}^{g} c_i^2 S_i^2 / N_i, \tag{3}$$

where $S_i^2 = \sum_{j=1}^{N_i} (Y_{ji} - \bar{Y}_i)^2 \Big/ (N_i - 1)$ is the sample variance for $i = 1, \ldots, g$. Then an approximate and useful pivotal quantity T for interval estimation of ψ can be expressed as

$$T = \frac{\widehat{\psi} - \psi}{\widehat{\Sigma}^{1/2}}. \tag{4}$$

Due to the dependence of $\widehat{\Sigma}$ on the sample variances $(S_1^2, \ldots, S_g^2)$, the exact distribution of T is fairly complicated. Accordingly, it is of practical interest to consider feasible approximations. Assume $X = \sum_{i=1}^{g} a_i X_i$, is a positive linear combination of independent chi-square variables, where a_i are positive constants, and

$X_i \sim \chi^2(f_i)$ are independent chi-square random variables with degrees of freedom f_i for $i = 1, \ldots, g$. Using a chi-square approximation to the distribution of X, Rossi (1975) and Welch (1947) showed that

$$X \dot{\sim} \frac{\xi}{\nu} \times \chi^2(\nu),$$

where $\xi = \sum_{i=1}^{g} a_i f_i$ and $\nu = \left\{ \sum_{i=1}^{g} a_i f_i \right\}^2 \Big/ \left\{ \sum_{i=1}^{g} a_i^2 f_i \right\}$. It is well known that the sample variances S_i^2 are distributed independently of each other and $(Ni - 1)S_i^2 / \sigma_i^2 \sim \chi^2\ (N_i - 1)$ for $i = 1, \ldots, g$. Hence, $\widehat{\Sigma}$ has the approximate distribution

$$\widehat{\Sigma} \dot{\sim} \frac{\Sigma}{\nu} \times \chi^2(\nu), \tag{5}$$

where $\Sigma = \sum_{i=1}^{g} c_i^2 \sigma_i^2 / N_i$ and $\nu = \left\{ \sum_{i=1}^{g} c_i^2 \sigma_i^2 / N_i \right\}^2 \Big/ \left\{ \sum_{i=1}^{g} c_i^4 \sigma_i^4 [N_i^2 (N_i - 1)] \right\}$. It readily follows from Equations 2 and 5 that the quantity T given in Equation 4 has a convenient approximate distribution

$$T \dot{\sim} t(\nu),$$

where $t(\nu)$ is a t distribution with degrees of freedom ν. For inferential purposes, the term of degrees of freedom ν is replaced by its counterpart $\hat{\nu}$ with direct substitution of $(S_1^2, \ldots, S_g^2)$ for $(\sigma_1^2, \ldots, \sigma_g^2)$ in ν, where

$$\hat{\nu} = \left\{ \sum_{i=1}^{g} c_i^2 S_i^2 / N_i \right\}^2 \Big/ \left\{ \sum_{i=1}^{g} c_i^4 S_i^4 / \left[N_i^2 (N_i - 1) \right] \right\}. \tag{6}$$

Thus, the adjustment gives the following modified distribution

$$T \dot{\sim} t(\hat{\nu}). \tag{7}$$

Accordingly, a 100 $(1 - \alpha)\%$ approximate two-sided confidence interval (L, U) for the contrast effect ψ can be constructed from Equation 7 where

$$L = \widehat{\psi} - t_{\hat{\nu},\, \alpha/2} \widehat{\Sigma}^{1/2}, \quad U = \widehat{\psi} + t_{\hat{\nu},\, \alpha/2} \widehat{\Sigma}^{1/2}$$

and $t_{\hat{\nu},\, \alpha/2}$ is the upper 100 $(\alpha/2)$ percentile of the t distribution $t(\hat{\nu})$. For ease of presentation, the half width of the 100 $(1 - \alpha)\%$ two-sided confidence interval (L, U) is denoted by

$$H = t_{\hat{\nu},\, \alpha/2} \widehat{\Sigma}^{1/2}.$$

It is clear that the actual half-width H depends on the confidence coefficient $1 - \alpha$, the sample sizes $(N_1, \ldots, N_g)$, and variance estimates $(S_1^2, \ldots, S_g^2)$.

For advance research design, it is desirable to determine the sample sizes required to achieve the designated precision properties of a confidence interval. Two useful principles concern the control of the expected half width and the tolerance probability of the half width within a preassigned value. Specifically, it is necessary to determine the required sample sizes such that the expected half width of a 100 $(1 - \alpha)\%$ confidence interval is within the given bound

$$E[H] \leq \delta, \tag{8}$$

where the expectation $E[H]$ is taken with respect to the joint distribution of $(S_1^2, \ldots, S_g^2)$, and δ (>0) is a constant. On the other hand, one may compute the sample sizes needed to guarantee, with a given tolerance probability, that the half width of a 100 $(1 - \alpha)\%$ confidence interval will not exceed the planned value

$$P\{H \leq \omega\} \geq 1 - \gamma, \tag{9}$$

where $1 - \gamma$ is the specified tolerance level and ω (>0) is a constant.

Given the involved property in the variance estimator $\widehat{\Sigma}$, it may be tempting to adapt a simplified approach to computing the expected half-width $E[H]$ and tolerance probability $P\{H \leq \omega\}$ by employing the approximate distribution of $\widehat{\Sigma}$ given in Equation 5 and the straightforward simplification of $t_{\hat{\nu},\,\alpha/2} \doteq t_{\nu,\,\alpha/2}$. However, an exact approach is considered here to provide more accurate results. For ease of explication, a detailed description of alternative formulation for H is presented in Appendix A. With the distributional properties presented in Appendix A for K, $\hat{\nu}$, and W, the evaluation of expected half-width $E[H]$ in Equation 8 can be simplified as

$$E[H] = E_K[K^{1/2}] \times E_B[t_{\hat{\nu},\,\alpha/2} W^{1/2}]. \tag{10}$$

The expectation $E_K[K^{1/2}]$ is taken with respect to the distribution of K, and it follows from the standard result of a chi-square distribution with $N_T - g$ degrees of freedom that $E_K[K^{1/2}] = 2^{1/2} \times \Gamma\{(N_T - g + 1)\}/\Gamma\{N_T - g/2\}$. On the other hand, the expectation $E_B[t_{\hat{\nu},\,\alpha/2} W^{1/2}]$ is taken with respect to the joint distribution of $(B_1, \ldots, B_{g-1})$ and does not permit a closed-form expression. Since the pseudo β random variable function is generally available in major statistical software packages, Monte Carlo integration approach is utilized to assess the actual value of $E_B[t_{\hat{\nu},\,\alpha/2} W^{1/2}]$. Similarly, for analytic clarity and computational ease, the probability $P\{H \leq \omega\}$ given in Equation 9 is expressed as

$$P\{H \leq \omega\} = E_B[F_K\{\omega^2/(t^2_{\hat{\nu},\,\alpha/2} W)\}], \tag{11}$$

where $F_K\{.\}$ is the cumulative density function of $K \sim \chi^2(N_T - g)$. Note that the cumulative density function of a chi-square distribution and pseudo beta random number generating function are readily available in standard software systems. As in the case of expected half width, Monte Carlo integration method is used

to perform the required computation of tolerance probability $P\{H \leq \omega\}$ in Equation 11 with the current computing capabilities.

As there may be several possible choices of sample sizes $(N_1, \ldots, N_g)$ that satisfy the chosen precision criterion in the process of sample size calculations, it is constructive to consider an appropriate design with a priori designated sample size ratios that leads to a unique and optimal result. For ease of illustration, the sample size ratios $(r_1, \ldots, r_g)$ are specified in advance with $r_i = N_i/N_1$, and consequently, the group allocation ratios $q_i = N_i/N_T = r_i/\sum_{j=1}^{g} r_j$ for $i = 1, \ldots, g$. Thus, the process is confined to deciding the minimum sample size N_1 (with $N_i = N_1 r_i,\ i = 2, \ldots, g$) required to achieve the selected precision level with the computational formulas of expected half width and tolerance probability in Equations 10 and 11, respectively. Specifically, the sample sizes $(N_{\mathrm{EW}_1}, \ldots, N_{\mathrm{EW}_g})$ needed for the expected half width of a 100 $(1 - \alpha)\%$ two-sided confidence interval (L, U) to fall within the designated bound δ are the minimum integers $(N_1, \ldots, N_g) = N_1(r_1, \ldots, r_g)$ such that $E[H] \leq \delta$. On the other hand, the sample sizes $(N_{\mathrm{TP}1}, \ldots, N_{\mathrm{TP}}g)$ required to guarantee with a given tolerance probability $(1 - \gamma)$ that the half width of a 100 $(1 - \alpha)\%$ two-sided confidence interval (L, U) will not exceed the planned range ω are the smallest integers $(N_1, \ldots, N_g) = N_1(r_1, \ldots, r_g)$ such that $P\{H \leq \omega\} \geq 1 - \gamma$.

The determinations of optimal sample sizes involve iterative algorithm not readily available in standard statistical packages and, therefore, require a special purpose computer program for performing the necessary computations. To enhance the applicability of these sample size techniques, supplementary SAS/IML (SAS Institute, 2011) computer programs are developed to perform the extensive calculations. Moreover, a detailed simulation study is performed next to evaluate the accuracy of the suggested sample size procedures under a variety of model configurations.

Empirical Assessments of Sample Size Calculations for Individual Comparisons

Due to the theoretical complications of the suggested methodology for precise interval estimation of the contrasts under heteroscedastic ANOVA settings, the features and performances of the sample size procedures need to be delineated and examined through numerical investigations. Explicitly, the empirical examination was conducted in two stages. The first stage presented sample size calculations for the two precision measures of expected half width and tolerance probability under several model configurations. Then, Monte Carlo simulation was performed to demonstrate the precision behavior for the recommended sample size formulas under the design characteristics specified in the first step.

Note that the determination of sample sizes needed for the chosen precision of the confidence interval procedure requires detailed specifications of the confidence level, magnitudes of variance components, contrast coefficients, and sample size ratios. For illustration, the bounds of the interval expected half-width criterion are chosen as $\delta = 1$ and 2, and the other precision assurance principle specified the tolerance probability and interval half bound as $1 - \gamma = 0.90$ and $\omega = 1$ and 2, respectively. The confidence level is fixed as $1 - \alpha = 0.95$ throughout this numerical study. Moreover, we focus on the situation of $g = 4$ with the heterogeneous variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_3^2) = (1, 4, 9, 16)$ and the contrast coefficients $(c_1, c_2, c_3, c_4) = (1, -1/3, -1/3, -1/3)$. To represent balanced and unbalanced patterns, three different settings of sample size ratios are considered: $(r_1, r_2, r_3, r_4) = (1, 2, 3, 4), (1, 1, 1, 1)$, and $(4, 3, 2, 1)$. The designated frameworks basically follow those in Jan and Shieh (2014) and Tomarken and Serlin (1986) with some modifications for the purpose of interval estimation rather than hypothesis testing. More important, the combined configurations were chosen to give a wide range of sample size settings so that they not only provide practically useful implications but also serve as a benchmark to demonstrate the robustness of the proposed sample size procedures. Accordingly, the necessary sample sizes $(N_{EW1}, N_{EW2}, N_{EW3}, N_{EW4})$ and $(N_{TP1}, N_{TP2}, N_{TP3}, N_{TP4})$ are computed with respect to the selected precision requirements of expected half width and of tolerance probability, respectively. The resulting sample sizes are presented in Table 1 for all six joint model configurations of two varying interval half bounds and three different sample size ratio settings.

In particular, when $\delta = \omega = 1$, the computed sample sizes under the expected half-width consideration are $(N_{EW1}, N_{EW2}, N_{EW3}, N_{EW4}) = (9, 18, 27, 36), (17, 17, 17, 17)$, and $(48, 36, 24, 12)$ for the three sample size ratio structures $(r_1, r_2, r_3, r_4) = (1, 2, 3, 4), (1, 1, 1, 1)$, and $(4, 3, 2, 1)$, respectively. Alternatively, the corresponding sample sizes associated with the tolerance probability criterion are $(N_{TP1}, N_{TP2}, N_{TP3}, N_{TP4}) = (12, 24, 36, 48), (21, 21, 21, 21)$, and $(64, 48, 32, 16)$ for the three sets of sample size ratios, respectively. From a practical standpoint, the total sample sizes, N_T, of the balanced structure are less than those of the unbalanced structure for both types of interval precisions. Conversely, the case with inverse pairing of heterogeneous variance $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 4, 9, 16)$ and sample size ratio $(r_1, r_2, r_3, r_4) = (4, 3, 2, 1)$ incurs the largest number of total sample size. As expected, the same phenomenon continues to exist for larger interval half bounds $\delta = \omega = 2$. However, the required sample sizes for $\delta = \omega = 2$ are comparatively smaller than those for $\delta = \omega = 1$ for both precision principles. Specifically, the reported sample sizes when $\delta = \omega = 2$ are $(N_{EW1}, N_{EW2}, N_{EW3}, N_{EW4}) = (4, 8, 12, 16), (5, 5, 5, 5)$, and $(16, 12, 8, 4)$, and $(N_{TP1}, N_{TP2}, N_{TP3}, N_{TP4}) = (5, 10, 15, 20), (7, 7, 7, 7)$, and $(24, 18, 12, 6)$ for the three distinct sample size ratio setups. Moreover, it is prudent to note that the two precision criteria impose unique and distinct precision characteristics on the confidence intervals of linear

TABLE 1

Computed Sample Size, Expected Half Width, and Tolerance Probability of the Proposed Approaches for 95% Two-Sided Confidence Interval of Contrast ψ With Interval Half-Bound $\delta = \omega = 1$ and 2, and Tolerance Probability $1 - \gamma = 0.90$, When $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 4, 9, 16)$ and $(c_1, c_2, c_3, c_4) = (1, -1/3, -1/3, -1/3)$

	Expected Half Width				Tolerance Probability			
Bound	Sample Sizes	Simulated $E[H]$	Attained $E[H]$	Relative Error (%)	Sample Sizes	Simulated $P\{H < \omega\}$	Attained $P\{H < \omega\}$	Relative Error (%)
$\delta = \omega = 1$	(9, 18, 27, 36)	0.9582	0.9573	0.0891	(12, 24, 36, 48)	0.9542	0.9539	0.0271
	(17, 17, 17, 17)	0.9968	0.9968	−0.0007	(21, 21, 21, 21)	0.9133	0.9125	0.0885
	(48, 36, 24, 12)	0.9625	0.9633	−0.0897	(64, 48, 32, 16)	0.9365	0.9377	−0.1257
$\delta = \omega = 2$	(3, 6, 9, 12)	1.9100	1.9074	0.1359	(5, 10, 15, 20)	0.9649	0.9706	−0.5928
	(5, 5, 5, 5)	1.9984	1.9967	0.0826	(7, 7, 7, 7)	0.9171	0.9199	−0.3090
	(16, 12, 8, 4)	1.9086	1.9102	−0.0834	(24, 18, 12, 6)	0.9251	0.9283	−0.3479

contrast and lead to fundamentally different magnitudes of desired sample sizes. According to the numerical assessment, it often requires a larger sample size to meet the necessary precision of tolerance probability than the control of a designated expected half width. The pattern of results between the two precision principles is similar to those reported in Kupper and Hafner (1989) and Shieh and Jan (2012). In the process of sample size calculations, the obtained precision levels associated with the reported sample sizes (N_{EW1}, N_{EW2}, N_{EW3}, N_{EW4}) and (N_{TP1}, N_{TP2}, N_{TP3}, N_{TP4}) should be less than or greater than the target value of interval half bound and tolerance probability, respectively. The actually achieved values of exact expected half-width $E[H]$ and tolerance probability $P\{H \leq \omega\}$ are also summarized in Table 1. The precision differences between the actual level and the nominal value are due to the underlying metric of integer sample sizes and the constraint of a designated sample size allocation ratio.

As mentioned earlier, one may attempt to simplify the distribution of interval half width as $H \dot{\sim} (t_{\nu,\,\alpha/2}\Sigma^{1/2}/\nu^{1/2})\{\chi^2(\nu)\}^{1/2}$, where ν is given in Equation 5. Consequently, the simple approximation gives

$$E[H] \doteq [t_{\nu,\,\alpha/2}\Sigma^{1/2}2^{1/2}\Gamma\{(\nu+1)/2\}]/[\nu^{1/2} \times \Gamma\{\nu/2\}], \tag{12}$$

and

$$P\{H < \omega\} \doteq F_{K^*}\left\{(\nu \times \omega^2)\Big/\left(t^2_{\nu,\,\alpha/2}\Sigma\right)\right\}, \tag{13}$$

where $F_{K^*}\{\cdot\}$ is the cumulative density function of $K^* \sim \chi^2(\nu)$. The two expressions in Equations 12 and 13 provide alternative formulas to compute the optimal sample sizes for precise interval estimation of contrast effects. For the prescribed design configurations along with three sample size ratio settings, the computed sample sizes are summarized in Table 2 where the resulting sample sizes (N_{EW1}, N_{EW2}, N_{EW3}, N_{EW4}) for the expected half-width consideration are (9, 18, 27, 36), (18, 18, 18, 18), and (44, 33, 22, 11) for $\delta = 1$, and (4, 8, 12, 16), (7, 7, 7, 7), and (20, 15, 10, 5) for $\delta = 2$. Also, the required sample sizes (N_{TP1}, N_{TP2}, N_{TP3}, N_{TP4}) associated with tolerance probability are (11, 22, 33, 44), (22, 22, 22, 22), and (56, 42, 28, 14) for $\delta = 1$, and (5, 10, 15, 20), (10, 10, 10, 10), and (24, 18, 12, 6) for $\delta = 2$. Although the simplified method gives the identical result with the proposed procedure for two of the six sets of sample sizes in both cases of precision criteria, the two formulations generally produce distinct behaviors and the calculated sample sizes can be substantially different in some cases. The attained precision levels of the approximate expected half width and approximate tolerance probability computed with Equations 12 and 13 are also presented in Table 2. More important, the resulting approximate precision outcomes differ from the exact expected half width and exact tolerance probability calculated with the recommended Equations 10 and 11, respectively. The adequacy and

TABLE 2

Computed Sample Size, Expected Half Width and Tolerance Probability of the Simplified Methods for 95% Two-Sided Confidence Interval of Contrast ψ With Interval Half-Bound $\delta = \omega = 1$ and 2, and Tolerance Probability $1 - \gamma = 0.90$, When $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1,\ 4,\ 9,\ 16)$ and $(c_1, c_2, c_3, c_4) = (1, -1/3, -1/3, -1/3)$

	Expected Half Width				Tolerance Probability			
Bound	Sample Sizes	Simulated $E[H]$	Attained $E[H]$	Relative Error (%)	Sample Sizes	Simulated $P\{H < \omega\}$	Attained $P\{H < \omega\}$	Relative Error (%)
$\delta = \omega = 1$	(8, 16, 24, 32)	1.0189	0.9846	3.3616	(9, 18, 27, 36)	0.6565	0.9620	−46.5323
	(17, 17, 17, 17)	0.9955	0.9790	1.6562	(18, 18, 18, 18)	0.6521	0.9437	−44.7155
	(44, 33, 22, 11)	1.0128	0.9719	4.0328	(48, 36, 24, 12)	0.6232	0.9396	−50.7727
$\delta = \omega = 2$	(3, 6, 9, 12)	1.9174	1.6736	12.7130	(4, 8, 12, 16)	0.8605	0.9999	−16.1962
	(5, 5, 5, 5)	2.0046	1.8581	7.3077	(6, 6, 6, 6)	0.7710	0.9655	−25.2324
	(16, 12, 8, 4)	1.9018	1.6870	11.2959	(20, 15, 10, 5)	0.8054	0.9967	−23.7511

discrepancy of the competing techniques are further evaluated through the following Monte Carlo simulation.

With the designated configurations and respective sample sizes for the exact and approximate methods listed in Tables 1 and 2, estimates of the true interval half width and tolerance probability are computed through Monte Carlo simulation of 10,000 independent data sets. For each replicate, the confidence limits and corresponding interval half width of the two-sided 95% confidence intervals of linear contrast are calculated. Then the simulated expected half width is the mean of the 10,000 replicates of interval half widths, whereas the simulated tolerance probability is the proportion of the 10,000 replicates whose values of interval half width are less than or equal to the specified bound. The adequacy of the sample size procedure for precise interval estimation is determined by one of the following formulas: relative error = (simulated expected half width − attained expected half width)/simulated expected half width or relative error = (simulated tolerance probability − attained tolerance probability)/simulated tolerance probability. Both the simulated values of expected half width and tolerance probability and the corresponding percentage of relative errors are summarized in Tables 1 and 2. According to the numerical results in Tables 1 and 2, the precision performance of the proposed sample size procedures maintains a close range near the nominal levels. Specifically, all the six absolute relative errors of the expected half width are less than 1%, and the absolute relative differences of tolerance probability have a maximum of 0.5928%. It can be seen that the performance of the proposed sample size procedures is fairly good for the range of model specifications considered here. However, the discrepancies between simulated half width and approximate half width in Table 2 indicate that the simplified method is not sufficiently accurate because the resulting relative errors range from 1.6562% to 12.7130%. Moreover, the other results of tolerance probability are not satisfactory either because the corresponding relative errors have a wide range of −16.1962% to −50.7727%. In view of these numerical evaluations, we conclude that the proposed procedures outperform the simplified methods in sample size calculations for precise interval estimation of contrast effects.

Multiple Comparisons of Means

The examination and methodology of individual comparisons of means are extended in this section to the general context of multiple comparisons involving a family of linear contrasts. Assume it is desirable to estimate L linear contrasts of the means denoted by

$$\psi_l = \sum_{i=1}^{g} c_{li}\mu_i,$$

and the corresponding unbiased estimator $\widehat{\psi}_l$ is given by

$$\widehat{\psi}_l = \sum_{i=1}^{g} c_{li}\overline{Y_i},$$

where c_{li} are the contrast coefficients with $\sum_{i=1}^{g} c_{li} = 0$ for $l = 1, \ldots, L$. In order to control the overall confidence level for the family of simultaneous confidence intervals $\{(L_l, U_l),\ l = 1, \ldots, L\}$, the construction of proper procedures is more involved than the situation of an individual interval estimation. Notably, several distinct and useful methods have been studied in Brown and Forsythe (1974), Dunnett (1980), Games and Howell (1976), Tamhane (1977), and Ury and Wiggins (1971). In particular, a total of six procedures are considered here for their unique feature and desirable property. The approach of Brown and Forsythe is applicable when the researcher is interested in multiple comparisons of complex linear contrasts, and the other five methods are only appropriate for multiple comparisons of pairwise mean differences.

The previous variance estimator and corresponding distribution of individual contrast are modified as $\widehat{\Sigma}_l = \sum_{i=1}^{g} c_{li}^2 S_i^2/N_i$, and

$$\widehat{\sum_{l}} \dot{\sim} \frac{\Sigma_l}{\nu_l} \times \chi^2(\nu_l),$$

where $\widehat{\Sigma}_l = \sum_{i=1}^{g} c_{li}^2\sigma_i^2/N_i$ and $\nu_l = \left\{\sum_{i=1}^{g} c_{li}^2\sigma_i^2/N_i\right\}^2 / \left\{\sum_{i=1}^{g} c_{li}^4\sigma_i^4/\left[N_i^2(N_i-1)\right]\right\}$ for $l = 1, \ldots, L$. The resulting confidence interval (L_l, U_l) of ψ_l is of the form

$$L_l = \widehat{\psi}_l - Q_{\hat{\nu}_l,\ \alpha}\widehat{\Sigma}_l^{1/2} \text{ and } U_l = \widehat{\psi}_l + Q_{\hat{\nu}_l,\ \alpha}\widehat{\Sigma}_l^{1/2},$$

where the critical value $Q_{\hat{\nu}l,\ \alpha}$ is suitably chosen to approximate the desired joint confidence level $1 - \alpha$, and the estimated degrees of freedom is

$$\hat{\nu}_l = \left\{\sum_{i=1}^{g} c_{li}^2 S_i^2/N_i\right\}^2 \Bigg/ \left\{\sum_{i=1}^{g} c_{li}^4 S_i^4/\left[N_i^2(N_i-1)\right]\right\},$$

for $l = 1, \ldots, L$. The half width of the two-sided confidence interval (L_l, U_l) is denoted by

$$H_l = Q_{\hat{\nu}_l,\ \alpha}\hat{\Sigma}_l^{1/2}.$$

In the particular case of pairwise multiple comparisons, the contrast coefficients are all zero except that $c_{li} = 1$ and $c_{li'} = -1, 1 \le i < i' \le g$, for $l = 1, \ldots, L = g(g-1)/2$. Thus, the expressions of $\widehat{\Sigma}_l$ and $\hat{\nu}_l$ can be specifically simplified as $\widehat{\Sigma}_l = \{S_i^2/N_i + S_{i'}^2/N_{i'}\}$ and $\hat{\nu}_l = \{S_i^2/N_i + S_{i'}^2/N_{i'}\}^2/\{S_i^4/[N_i^2(N_i-1)] + S_{i'}^4/[N_{i'}^2(N_{i'}-1)]\}$, respectively.

The following six multiple comparison procedures for choosing $Q_{\hat{\nu}_l,\ \alpha}$ are considered:

1. The approximate confidence intervals of all contrasts proposed in Brown and Forsythe (1974) have

$$Q_{\hat{\nu}_l},\ \alpha = \{(g-1)F_{g-1,\hat{\nu}_{l,\,\alpha}}\}^{1/2}, \tag{14}$$

where $F_{g-1,\hat{\nu}_{l,\,\alpha}}$ is the upper α quantile of an F distribution with degrees of freedom $g-1$ and $\hat{\nu}_l$.

2. The approximate confidence intervals for all pairwise differences suggested in Ury and Wiggins (1971) use

$$Q_{\hat{\nu}_l},\ \alpha = t_{\hat{\nu}_{l,\,\alpha B}} \text{ and } \alpha_B = \alpha/(2L). \tag{15}$$

3. Games and Howell (1976) proposed the following expression

$$Q_{\hat{\nu}_{l,\,\alpha}} = q_{g,\,\hat{\nu}_{l,\,\alpha}}/2^{1/2}, \tag{16}$$

where $q_{g\,\hat{\nu}_l,\alpha}$ denotes the upper α quantile of the studentized range distribution with parameters g and $\hat{\nu}_l$

4. Tamhane (1977) employed the critical value

$$Q_{\hat{\nu}_{l,\,\alpha}} = t_{\hat{\nu}_{l,\,\alpha_T}} \text{ and } \alpha_T = \left\{1-(1-\alpha)^{1/L}\right\}\Big/2. \tag{17}$$

5. Dunnett (1980) modified the notion of Cochran (1964) and suggested the designated quantity

$$Q_{\hat{\nu}_{l,\,\alpha}} = q^*_{g,\hat{\nu}_{l,\,\alpha}}/2^{1/2}, \tag{18}$$

where $q^*_{g,\hat{\nu}_{l,\,\alpha}} = \{q_{g,\,N_i-1,\,\alpha}S_i^2/N_i + q_{g,\,N_i-1,\,\alpha}S_{i'}^2/N_{i'}\}/\{S_i^2/N_i + S_{i'}^2/N_{i'}\}$.

6. Dunnett (1980) also considered the alternative form

$$Q_{\hat{\nu}_{l,\,\alpha}} = m_{L,\,\hat{\nu}_{l,\,\alpha}}, \tag{19}$$

where $m_{L,\,\hat{\nu}_{l,\,\alpha}}$ denotes the upper α quantile of the studentized maximum modulus distribution with parameters L and $\hat{\nu}_l$.

To enhance the applicability of these multiple comparison procedures, we apply the expected half width and tolerance probability criteria to determine the required sample sizes for precise interval estimation of both general and pairwise contrasts. It is essential to note that the aforementioned critical values given in Equations 14 through 19 differ among all contrasts. However, Pan and Kupper (1999) focused on the multiple comparison methods under the homogeneous variance and balanced design, and the corresponding critical values remain identical for the whole family of contrasts. Consequently, the underlying properties of the resulting interval half widths in Pan and Kupper (1999) are relatively simpler than those within the context of heteroscedastic and unbalanced settings. Because of the complex nature of the interval half widths, complete analytical assessments of the joint properties of expected half width and tolerance probability are

not feasible. Alternative arguments and simplifications are developed to circumvent theoretical difficulties and permit useful applications.

In view of the critical implications of desired precision for the resulting confidence intervals, the notion of expected half width for sample size calculations is frequently introduced in standard texts. For multiple comparisons, it is necessary to determine the required sample sizes $\{N_{\mathrm{EW}1}, \ldots, N_{\mathrm{EW}g}\}$ such that all the expected half widths of the simultaneous confidence intervals $\{(L_l, U_l), l = 1, \ldots, L\}$ for contrast effects are within the given bound $E[H_l] \leq \delta_l$, where δ_l (>0) are the designated bounds for $l = 1, \ldots, L$. Then it is equivalent to consider that

$$\max_{1 \leq l \leq L} \{E[H_l/\delta_l]\} \leq 1. \tag{20}$$

Unfortunately, the appraisal of expected half-width $E[H_l/\delta_l]$ cannot be solved analytically. It is shown in the Appendix B that the condition given in Equation 20 can be alternatively evaluated by

$$E[H_{l^*}] \leq \delta_{l^*}, \tag{21}$$

where l^* is the value l so that $\Delta_l = N_1^{1/2} \sum_l^{1/2} / \delta_l$ attains the maximum for $l = 1, \ldots, L$. Therefore, the alternative condition given in Equation 21 permits an enormous simplification of the joint appraisal of several bounded interval half widths and the previous approach to compute the exact expected half-width $E[H]$ in Equation 10 of individual confidence interval can be readily applied to calculate the exact value $E[H_{l^*}]$. More importantly, with the initially specified sample size ratios $(r_1, \ldots, r_g)$, the sample sizes $(N_{\mathrm{EW}1}, \ldots, N_{\mathrm{EW}g}) = N_1(r_1, \ldots, r_g)$ required to ensure $E[H_l] \leq \delta_l$, $l = 1, \ldots, L$, can be determined by the minimum integer N_1 such that $E[H_{l^*}] \leq \delta_{l^*}$.

In addition, the criterion of tolerance probability of interval half width within a given value is of special interest. Hence, it is desirable to find the sample sizes $\{N_{\mathrm{TP}1}, \ldots, N_{\mathrm{TP}g}\}$ needed to guarantee, with a given joint tolerance probability $1 - \gamma$, that the half widths of the simultaneous confidence intervals $\{(L_l, U_l), l = 1, \ldots, L\}$ for contrast effects will not exceed the planned range

$$P\{H_l \leq \omega_l, \ l = 1, \ \ldots, \ L\} \geq 1 - \gamma, \tag{22}$$

where ω_l (> 0) are the designated bounds for $l = 1, \ldots, L$. To provide a feasible solution for the joint tolerance probability, an approximate expression is shown in Appendix C for the condition given in Equation 22

$$P\{H_{l^*} \leq \omega_{l^*}\} \geq 1 - \gamma, \tag{23}$$

where l^* is the value l so that $\Omega_l = N_1^{1/2} \sum_l^{1/2} / \omega_l$ attains the maximum for $l = 1, \ldots, L$. Notably, the alternative formulation in Equation 23 affords a major simplification of the combined evaluation of several interval half widths and the

prescribed method to compute the exact tolerance probability $P\{H \leq \omega\}$ in Equation 11 of individual confidence interval can be immediately applied to compute the exact value $P\{H_{l^*} \leq \omega_{l^*}\}$. In short, for the formerly designated sample size ratios $(r_1, \ldots, r_g)$, the sample sizes $(N_{\mathrm{TP}1}, \ldots, N_{\mathrm{TP}g}) = N_{\mathrm{TP}1}(r_1, \ldots, r_g)$ required to ensure the joint tolerance probability $P\{H_l \leq \omega_l, \ l = 1, \ldots, L\} \geq 1 - \gamma$, is computed by the minimum integer N_1 so that $P\{H_{l^*} \leq \omega_{l^*}\} \geq 1 - \gamma$.

For the special case of multiple comparisons pertaining to the pairwise contrasts of each different treatment to a control, the three approaches of Ury and Wiggins (1971), Tamhane (1977), and Dunnett (1980) given in Equations 15, 17, and 19 can be readily modified with $L = g - 1$ to construct the simultaneous confidence intervals. The corresponding sample size determinations can also be conducted with the suggested methods. For practical applications, computer algorithms are required for performing the sample size calculations so that the simultaneous confidence intervals of the six multiple comparison procedures will attain the desired precision. Empirical illustrations are presented next to demonstrate the usefulness and accuracy of the proposed sample size procedures and supplementary SAS/IML (SAS Institute, 2011) computing algorithms.

Empirical Assessments of Sample Size Calculations for Multiple Comparisons

The similarities and differences among the proposed sample size procedures for multiple comparisons are demonstrated here through the model formulations in the previous illustration of individual comparisons. In this case, we address the sample size problem for the family of simultaneous confidence intervals for pairwise multiple comparisons. A systematic numerical investigation of four-group heteroscedastic ANOVA is conducted by fixing the confidence level $1 - \alpha = 0.95$ and heterogeneous error variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 4, 9, 16)$, and varying the sample size allocation ratio: (1, 2, 3, 4), (1, 1, 1, 1), and (4, 3, 2, 1). Similar to the implementation of the preceding examination, this empirical study includes sample size calculation and Monte Carlo simulation.

For the designated multiple comparison procedure to ensure all six two-sided confidence intervals of pairwise mean differences have expected half widths within the bound $\delta_l = \delta = 2$, the necessary sample sizes ($N_{\mathrm{EW}1}$, $N_{\mathrm{EW}2}$, $N_{\mathrm{EW}3}$, $N_{\mathrm{EW}4}$) computed with suggested technique are summarized in Table 3. Moreover, the sample sizes $(N_{\mathrm{TP}1}, \ldots, N_{\mathrm{TP}4})$ are also presented when it is required to guarantee, with a given tolerance probability $1 - \gamma = 0.90$, that the half widths of all six two-sided confidence intervals for pairwise mean differences will not exceed the planned range $\omega_l = \omega = 2$. It should be clear from the optimal sample sizes presented in Table 3 that the required sample sizes under the tolerance probability consideration are still larger than those for the expected half-width criterion for all six multiple comparison procedures. Also, it is interesting to note that the computed sample sizes for the three methods of Ury and Wiggins (1971), Tamhane

TABLE 3

Computed Sample Size, Expected Half Width and Tolerance Probability of the Proposed Approaches for Simultaneous 95% Two-Sided Confidence Intervals of Pairwise Contrasts With Interval Half-bound $\delta = \omega = 2$ and Tolerance Probability $1 - \gamma = 0.90$, When $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1,\ 4,\ 9,\ 16)$

	Expected Half Width				Tolerance probability			
Procedure	Sample Sizes	Simulated Maximum $E[H]$	Approximate Maximum $E[H]$	Relative Error (%)	Sample Sizes	Simulated Joint $P\{H < \omega\}$	Approximate Joint $P\{H < \omega\}$	Relative Error (%)
Brown and Forsythe	(15, 30, 45, 60)	1.9361	1.9367	−0.0309	(17, 34, 51, 68)	0.9318	0.9367	−0.5211
	(51, 51, 51, 51)	1.9894	1.9880	0.0691	(60, 60, 60, 60)	0.9203	0.9126	0.8386
	(168, 126, 84, 42)	1.9975	1.9957	0.0876	(204, 153, 102, 51)	0.9041	0.9066	−0.2801
Ury and Wiggins	(13, 26, 39, 52)	1.9762	1.9769	−0.0387	(15, 30, 45, 60)	0.9002	0.9020	−0.2022
	(46, 46, 46, 46)	1.9871	1.9860	0.0516	(54, 54, 54, 54)	0.8984	0.9028	−0.4856
	(152, 114, 76, 38)	1.9977	1.9939	0.1921	(188, 141, 94, 47)	0.9202	0.9204	−0.0199
Game and Howell	(12, 24, 36, 48)	1.9973	1.9980	−0.0333	(15, 30, 45, 60)	0.9597	0.9608	−0.1157
	(43, 43, 43, 43)	1.9980	1.9942	0.1921	(52, 52, 52, 52)	0.9264	0.9246	0.1961
	(144, 108, 72, 36)	1.9878	1.9870	0.0399	(176, 132, 88, 44)	0.9032	0.9067	−0.3841
Tamhane	(13, 26, 39, 52)	1.9706	1.9713	−0.0387	(15, 30, 45, 60)	0.9090	0.9093	−0.0303
	(46, 46, 46, 46)	1.9814	1.9804	0.0516	(54, 54, 54, 54)	0.9067	0.9097	−0.3316
	(152, 114, 76, 38)	1.9919	1.9881	0.1921	(188, 141, 94, 47)	0.9258	0.9256	0.0240
Dunnett and Cochran	(12, 24, 36, 48)	1.9509	1.9525	−0.0830	(14, 28, 42, 56)	0.9232	0.9270	−0.4151
	(45, 45, 45, 45)	1.9836	1.9843	−0.0346	(53, 53, 53, 53)	0.9094	0.9157	−0.6902
	(152, 114, 76, 38)	1.9854	1.9845	0.0478	(184, 138, 92, 46)	0.9218	0.9183	0.3748
Dunnett	(13, 26, 39, 52)	1.9684	1.9682	0.0089	(15, 30, 45, 60)	0.9113	0.9127	−0.1460
	(46, 46, 46, 46)	1.9791	1.9774	0.0869	(54, 54, 54, 54)	0.9098	0.9136	−0.4124
	(152, 114, 76, 38)	1.9885	1.9848	0.1858	(184, 138, 92, 46)	0.9037	0.9074	−0.4074

(1977), and Dunnett (1980) given in Equations 15, 17, and 19 provide almost the identical results for all six combined cases of different sample size ratio and precision principle. Also, Brown and Forsythe's (1974) procedure appears to require the largest sample sizes, while the method of Games and Howell (1976) tends to give the least sample sizes for the model configurations considered here. Also, the sample sizes associated with the Dunnett's (1980) modified approach of Cochran (1964) are slightly smaller than those of the three procedures of Ury and Wiggins (1971), Tamhane (1977), and Dunnett (1980).

For ease of exposition, Table 3 specifically shows the corresponding approximate maximum expected half width and approximate joint tolerance probability for all design settings. These values are compared with the respective simulated maximum expected half width and simulated joint tolerance probability obtained from Monte Carlo simulation. With the design configurations and respective sample sizes for the multiple comparison methods, the simulated maximum expected half width is the largest value of the six means of the 10,000 replicates of interval half widths, whereas the simulated joint tolerance probability is the proportion of the 10,000 replicates whose values of all six interval half width are less than or equal to the specified bound. Consequently, the adequacy of the suggested sample size procedures for precise interval estimation is determined by the discrepancy between the nominal levels of simulated maximum half width and approximate maximum half width, or the difference between the simulated joint tolerance probability and approximate joint tolerance probability. Accordingly, the simulated results and corresponding relative errors listed in Table 3 clearly show that the proposed sample size formulas perform extremely well because all absolute relative errors are less than 0.01 for the 36 cases examined here.

Note that the computations of the quantiles of the studentized range distribution and the studentized maximum modulus distribution require a special function such as the SAS PROBMC function which may not be readily available in other software systems. To ease the burden of the extensive and iterative process in sample size determinations, the computations of quantile values $q_{g,\hat{\nu}_l,\alpha}, q^*_{g,\hat{\nu}_l,\alpha}$, and $m_{L,\hat{\nu}_l,\alpha}$ in Equations 16, 18, and 19, respectively, can be simplified by using the respective values $q_{g,\hat{\nu}_l,\alpha}, q^*_{g,\hat{\nu}_l,\alpha}$, and $m_{L,\hat{\nu}_l,\alpha}$ with the degrees of freedom $\hat{\nu}$ being replaced by its parameter counterpart ν. As explained in the theoretical derivations, the discrepancy should be negligible for moderately large degrees of freedom. More important, our numerical results demonstrate the modification leads to a substantially more efficient algorithm and also maintain sensible accuracy. It is prudent to examine the behavior of the suggested techniques in a variety of other situations. However, these empirical evidences demonstrate that the proposed sample size procedures provide feasible and accurate solutions to precise simultaneous confidence interval estimation of the six multiple comparison procedures under a wide variety of heteroscedastic model configurations.

Numerical Illustrations

To demonstrate the features of the suggested procedures in sample size planning, the National Assessment Educational Progress educational achievement data considered in Williams, Jones, and Tukey (1999) is used as an example. Specifically, the tabulated values shown in their Table 3 represent the means and standard errors of the eighth-grade mathematics proficiency changes between 1990 and 1992 for the 34 states. However, unlike the demonstration of alternative hypothesis testing procedures in Williams et al., we focus on the sample size calculations for interval estimation of the differences in achievement changes between the eight states in the northeast region. First, the individual comparison may take up the scenario to compare the outcomes of New Jersey and average of the other seven states in the region. To illustrate sample size determination for design planning, the reported summary statistics are modified as population mean change and standard deviation parameters. Because the sample standard deviations are not available, for the sake of explication, the standard deviations are set as 5 times of the sample standard errors. In the following sample size calculations, the mean changes and associated standard deviations are $\mu = (1.565, 1.374, 3.399, 4.893, 4.303, 3.204, 4.422, 5.097)$ and $\sigma = 5 \times (1.927, 1.347, 1.923, 2.532, 2.205, 1.534, 1.354, 0.948)$ for the states of New Jersey, Delaware, Maryland, New York, Pennsylvania, Connecticut, New Hampshire, and Rhode Island, respectively. With the additional settings of confidence coefficient $1 - \alpha = 0.95$, interval half-width bounds $\delta = \omega = 2.5$, tolerance level $1 - \gamma = 0.90$, the computed sample sizes for balanced design are 66 and 78 for each group under the expected half width and tolerance probability criterion, respectively. The actual specifications of these configurations are incorporated in the SAS/IML programs presented as the supplementary files.

Second, to ensure all pairwise confidence intervals between the achievement changes of the eight states are narrow enough to yield meaningful precision, the necessary sample sizes can be calculated with the developed algorithms for the six different multiple comparison procedures described earlier. Using the previously mentioned model configurations, the required sample sizes to meet the desired expected half width with pairwise difference $\delta_l = 2.5$, $l = 1, \ldots, 7$, were calculated with the supplementary SAS/IML programs. Accordingly, the resulting sample sizes for the six procedures of Brown and Forsythe (1974), Ury and Wiggins (1971), Games and Howell (1976), Tamhane (1977), and Dunnett (1980) are 637, 443, 417, 441, 419, and 441, respectively. On the other hand, the corresponding sample sizes to guarantee the joint tolerance probability is at least $1 - \gamma = 0.90$ with the desired half-widths $\omega_l = 2.5, l = 1, \ldots, 7$, are 669, 470, 442, 468, 444, and 467 for the six multiple comparison methods. Clearly, the required sample sizes are substantially larger than those in the individual comparisons. With these numerical illustrations, users can easily identify

the statements containing the key values in the computer code and then modify the program to accommodate their own model specifications.

Conclusion

The editorial policies and statistical guidelines of several prominent educational and psychological journals called for greater use of confidence intervals for principal effect sizes. Accordingly, it has become consensus across many scientific disciplines to include appropriate effect size measures and associated confidence intervals when documenting the results of research studies. From a study-planning point of view, researchers may wish to credibly address specific research questions and confirm meaningful treatment effects, so that the resulting confidence interval will meet the designated precision requirements. The general formulation of a linear combination of population means permits a wide range of research questions to be evaluated within the context of ANOVA. Accordingly, a linear contrast between two or more means represents an effect size index in the individual and multiple comparison investigations.

In order to enhance the applicability of single and simultaneous confidence intervals within the framework of one-way heteroscedastic ANOVA, this study presents the corresponding sample size techniques under two precision principles. The precision criteria consist of the control of the expected width and the assurance of tolerance probability of confidence intervals. It is noteworthy that the two principles of expected width and tolerance probability are closely related to the two standard criteria of unbiasedness and consistency in statistical point estimation, respectively. In other words, these two measures impose unique and distinct aspects of precision characteristics on the resulting confidence intervals, and each principle has conceptual and empirical implications in its own right. For most of the situations, prior knowledge or theory alone enables us to determine the appropriate magnitude of interval half width because its scale is the same as that of the linear contrast. On the other hand, the suitable values of tolerance levels are within the range of 0.70 to 0.99 as demonstrated in Kupper and Hafner (1989).

Consequently, the suggested sample size procedures update and expand upon current work of Pan and Kupper (1999) and related results in the literature. Although the discussion concentrated on the one-way ANOVA setting, the principles and procedures are also applicable in more complicated factorial and extended formulations. Detailed sample size tables are presented to help researchers have a better understanding of the intrinsic relationships that exists between the optimal sample sizes, model characteristics, and precision considerations. Since existing software packages do not accommodate sample size calculations with the same degree of generality as illustrated in this research, computer programs are also developed to aid the use of the suggested procedures. The proposed sample size methodology should be useful for practical purposes of planning individual and multiple comparison studies in which variances differ across groups.

Appendix A

Alternative Formulation of Interval Half Width

In order to conduct exact and efficient computations, the following alternative formulation for H is derived from the expression of $\hat{\Sigma}$ given in Equation 3

$$H = t_{\hat{\nu},\,\alpha/2}\{K \times W\}^{1/2},$$

where $K = \sum_{i=1}^{g} K_i \sim \chi^2(N_T - g), K_i \sim \chi^2(N_i - 1), N_T = \sum_{i=1}^{g} N_i$ $W = \sum_{i=1}^{g} b_i A_i$, $b_i = (c_i^2\sigma_i^2)/\{N_i(N_i-1)\}$ and $A_i = K_i/K$, $i = 1, \ldots, g$. Note that the approximate degrees of freedom $\hat{\nu}$ given in Equation 6 can also be expressed as $\hat{\nu} = \left\{\sum_{i=1}^{g} b_i A_i\right\}^2 \Big/ \left\{\sum_{i=1}^{g} b_i^2 A_i^2/(N_i - 1)\right\}$. Moreover, it is computationally simple and relatively stable to rewrite the dependence of $(A_1, \ldots, A_g)$ on the chi-square random variables in terms of the beta random variables, see Johnson, Kotz, and Balakrishnan (1995, p. 212). Specifically, $A_1 = \prod_{i=1}^{g-1} B_i, A_2 = (1 - B_1)\prod_{i=2}^{g-1} B_i, \ldots, A_{g-1} = (1 - B_{g-2})B_{g-1}$, and $A_g = 1 - B_{g-1}$, where $B_i = \left\{\sum_{j=1}^{i} K_j\right\} \Big/ \left\{\sum_{j'=1}^{i+1} K_{j'}\right\}$ has a beta distribution with $B_i \sim \text{beta}\left\{\sum_{j=1}^{i}(N_j - 1)/2,\ (N_{i+1} - 1)/2\right\}$ for $i = 1, \ldots, g-1$. An important underlying property of the suggested formulations is that the random variables $B_1, \ldots, B_{g-1}$ and K are mutually independent. Hence, both $\hat{\nu}$ and W can be viewed as a function of beta random variables $(B_1, \ldots, B_{g-1})$, and they are independent of K.

Appendix B

Approximate Expression of Equation 20

Consider the approximate evaluation for the expected half-width $E[H_l/\delta_l]$

$$E[H_l/\delta_l] \doteq E[(Q_{\nu l,\,\alpha}\Sigma_l^{1/2}/\delta_l)(U_l/\nu_l^{1/2})] \doteq Q_{\nu l,\,\alpha}\Sigma_l^{1/2}/\delta_l,$$

where $U_l \sim \chi^2(\nu_l)$ for $l = 1, \ldots, L$. Assume the maximum of $\Delta_l = N_1^{1/2}\sum_l^{1/2}/\delta_l = \left(\sum_{i=1}^{g} c_{li}^2\sigma_i^2/r_i\right)^{1/2}/\delta_l$, $l = 1, \ldots, L$, occurs when $l = l^*$ with

$$\Delta_{l^*} = \max_{1 \le l \le L} \Delta_l = \left(\sum_{i=1}^{g} c_{l^*i}^2\sigma_i^2/r_i\right)^{1/2}/\delta_{l^*}.$$

Note that the critical values $Q_{\nu l,\,\alpha}$ depend on the sample sizes $\{N_1, \ldots, N_g\}$ and are not substantially different from each other for moderately large degrees of freedom ν_l. Essentially, the dominant term in $Q_{\nu l,\,\alpha}\Sigma_l^{1/2}/\delta_l$ is $\Sigma_l^{1/2}/\delta_l = \Delta_l/N_1^{1/2}$. Hence, $\max_{1\le l\le L}\{E[H_l/\delta_l]\} \doteq \max_{1\le l\le L}\{Q_{\nu l,\,\alpha}\Sigma_l^{1/2}/\delta_l\} = \max_{1\le l\le L}\{Q_{\nu l,\,\alpha}\Delta_l/N_1^{1/2}\} \doteq Q_{\nu l^*,\,\alpha}\Delta_{l^*}/N_1^{1/2} = Q_{\nu l^*,\,\alpha}\Sigma_{l^*}^{1/2}/\delta_{l^*}$. The result implies that $\max_{1\le l\le L}\{E[H_l\delta_l]\} \doteq E[H_{l^*}/\delta_{l^*}]$, and the condition given in Equation 20 can be alternatively evaluated by

$$E[H_{l^*}] \le \delta_{l^*}.$$

Appendix C

Approximate Expression of Equation 22

Note that an useful approximate formulation of $P\{H_l \le \omega_l\}$ is

$$P\{H_l \le \omega_l\} = P\{H_l/\omega_l \le 1\} \doteq P\left\{\left(Q_{\nu_l,\,\alpha}\Sigma_l^{1/2}/\omega_l\right)(U_l/\nu_l)^{1/2} \le 1\right\},$$

where $U_l \sim \chi^2(\nu_l)$ for $l = 1, \ldots, L$. Assume the maximum of $\Omega_l = N_1^{1/2}\Sigma_l^{1/2}/\omega_l = \left(\sum_{i=1}^{g} c_{li}^2\sigma_i^2/r_i\right)^{1/2}/\omega_l$ occurs when $l = l^*$ with

$$\Omega_{l^*} = \max_{1\le l\le L}\Omega_l = \left(\sum_{i=1}^{g} c_{l^*i}^2\sigma_i^2/r_i\right)^{1/2}\Big/\omega_{l^*}.$$

Also, $P\{U_l/\nu_l \le a\}$ are fairly equivalent when the degrees of freedom ν_l are moderately large and the constant a is substantially greater than 1. In addition, the dominant term in $Q_{\nu l,\alpha}\,(\Sigma_l^{1/2}/\omega_l)$ is $\Sigma_l^{1/2}/\omega_l = \Omega_l/N_1^{1/2}$ because the critical values $Q_{\nu l,\,\alpha}$ are relatively close to each other in magnitude for moderately large degrees of freedom ν_l. Therefore, $P\{H_l \le \omega_{l,}\ l = 1, \ldots, L\} = P\{\max_{1\le l\le L}(H_l\omega_l) \le 1\} \doteq P\left\{\max_{1\le l\le L}(Q_{\nu l,\,\alpha}\Sigma_l^{1/2}/\omega_l)(U_l/\nu_l)^{1/2} \le 1\right\} \doteq P\{Q_{\nu l^*,\alpha}\,(\Omega_{l^*}/N_1^{1/2})(U_{l^*}/\nu_{l^*})^{1/2} \le 1\} = P\{Q_{\nu_{l^*},\,\alpha}(\Sigma_{l^*}^{1/2}/\omega_{l^*})(U_{l^*}/\nu_{l^*})^{1/2} \le 1\} \doteq P\{Q_{\nu_{l^*},\alpha}\,\widehat{\Sigma}_{l^*}^{1/2}/\omega_{l^*} \le 1\}$. Accordingly, $P\{H_l \le \omega_{l,}\ l = 1, \ldots, L\} \doteq P\{H_{l^*} \le \omega_{l^*}\}$ and the condition given in Equation 22 can be approximately evaluated as

$$P\{H_{l^*} \le \omega_{l^*}\} \ge 1 - \gamma.$$

Acknowledgment

The authors would like to thank the editor, Dr. Sandip Sinharay, and four anonymous reviewers for constructive suggestions that led to improved presentation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research of the first author was partially supported by National Science Council Grant NSC-102-2118-M-033-001.

References

Alhija, F. N. A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement*, *69*, 245–265.

American Educational Research Association Task Force on Reporting of Research Methods. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Research*, *35*, 33–40.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement*, *62*, 197–226.

Bird, K. D. (2004). *Analysis of variance via confidence intervals*. London, England: Sage.

Bretz, F., Hothorn, T., & Westfall, P. H. (2010). *Multiple comparisons using R*. Boca Raton, FL: Chapman & Hall/CRC.

Brown, M. B., & Forsythe, A. B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, *30*, 719–724.

Cochran, W. G. (1964). Approximate significance levels of the Behrens-Fisher test. *Biometrics*, *20*, 191–195.

Cohen, J. (1990). Things I have learned so far. *American Psychologist*, *45*, 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge/Taylor & Francis.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association*, *75*, 796–800.

Dunst, C. J., & Hamby, D. W. (2012). Guide for calculating and interpreting effect sizes and confidence intervals in intellectual and developmental disability research studies. *Journal of Intellectual & Developmental Disability*, *37*, 89–99.

Fenstad, G. U. (1983). A comparison between the *U* and *V* tests in the Behrens-Fisher problem. *Biometrika*, *70*, 300–302.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2–18.

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal *N*'s and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, *1*, 113–125.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*, 155–165.

Hahn, G. J., & Meeker, W. Q. (1991). *Statistical intervals: A guide for practitioners*. New York, NY: John Wiley.

Hochberg, Y., & Tamhane, A. J. (1987). *Multiple comparison procedures*. New York, NY: John Wiley.

Hsu, J. (1996). *Multiple comparisons: Theory and methods*. London, England: Chapman and Hall.

Jan, S. L., & Shieh, G. (2014). Sample size determinations for Welch's test in one-way heteroscedastic ANOVA. *British Journal of Mathematical and Statistical Psychology, 67*, 72–93.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York, NY: Wiley.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation and the Health Professions*, *26*, 258–287.

Kramer, C. (1956). Extension of multiple range tests to groups means with unequal numbers of replications. *Biometrics*, *12*, 307–310.

Kupper, L. L., & Hafner, K. B. (1989). How appropriate are popular sample size formulas? *The American Statistician*, *43*, 101–105.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw-Hill.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*, 188–209.

Liu, X. S. (2009). Sample size and the width of the confidence interval for mean difference. *British Journal of Mathematical and Statistical Psychology*, *62*, 201–215.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Odgaard, E. C., & Fowler, R. L. (2010). Confidence intervals for effect sizes: Compliance and clinical significance in the journal of consulting and clinical psychology. *Journal of Consulting and Clinical Psychology*, *78*, 287–297.

Pan, Z., & Kupper, L. L. (1999). Sample size determination for multiple comparison studies treating confidence interval width as random. *Statistics in Medicine*, *18*, 1475–1488.

Robey, R. R. (2004). Reporting point and interval estimates of effect-size for planned contrasts: Fixed within effect analyses of variance. *Journal of Fluency Disorders*, *29*, 307–341.

Rossi, J. A. D. (1975). An application of Welch's approximate *t*-solution of the Behrens-Fisher problem to confidence intervals. *Technometrics*, *17*, 57–60.

SAS Institute. (2011). SAS/IML User's Guide (Version 9.2) [Computer software]. Cary, NC: SAS Institute.

Satterthwaite, F. E. (1946). An approximate distribution of estimate of variance components. *Biometrics Bulletin*, *2*, 110–114.

Scheffe, H. (1959). *The analysis of variance*. New York, NY: Wiley.

Shieh, G., & Jan, S. L. (2012). Optimal sample sizes for precise interval estimation of Welch's procedure under various allocation and cost considerations. *Behavior Research Methods*, *44*, 202–212.

Smith, H. F. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research*, *9*, 211–212.

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*, 989–1004.

Tamhane, A. C. (1977). Multiple comparisons in model I one-way ANOVA with unequal variances. *Communications in Statistics*, *A6*, 15–32.

Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association*, *74*, 471–480.

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90–99.

Tukey, J. W. (1994). *The problem of multiple comparisons.* In H. Braun (Ed.), *The collected works of John W. Tukey VIII. Multiple comparisons: 1948–1983* (pp. 1–300). New York: Chapman and Hall.

Ury, H. K., & Wiggins, A. D. (1971). Large sample and other multiple comparisons among means. *British Journal of Mathematical and Statistical Psychology*, *24*, 174–194.

Wang, Y., & Kupper, L. L. (1997). Optimal sample sizes for estimating the difference in means between two normal populations treating confidence interval length as a random variable. *Commemorations in Statistics-Theory and Methods*, *26*, 727–741.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362.

Welch, B. L. (1947). The generalization of students' problem when several different population variances are involved. *Biometrika*, *34*, 28–35.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (2011). *Multiple comparisons and multiple tests: Using the SAS system* (2nd ed.). Cary, NC: SAS Institute.

Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, *38*, 29–60.

Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, *24*, 42–69.

Authors

SHOW-LI JAN is a professor of applied mathematics, Chung Yuan Christian University, Chungli, Taiwan; e-mail: sljan@math.cycu.edu.tw. Her research focuses on nonparametric methods and multiple test procedures.

GWOWEN SHIEH is a professor of management science, National Chiao Tung University, Hsinchu, Taiwan; e-mail: gwshieh@mail.nctu.edu.tw. His current research interests include sample size methodology and research methods.

Manuscript received June 04, 2013
Revision received October 19, 2013
Accepted January 09, 2014