PLOS ONE

RESEARCH ARTICLE

Optimal contrast analysis with heterogeneous variances and budget concerns

Show-Li Jan[1], **Gwowen Shieh**[2]*

1 Department of Applied Mathematics, Chung Yuan Christian University, Taiwan, Republic of China,
2 Department of Management Science, National Chiao Tung University, Taiwan, Republic of China

* gwshieh@mail.nctu.edu.tw

OPEN ACCESS

Citation: Jan S-L, Shieh G (2019) Optimal contrast analysis with heterogeneous variances and budget concerns. PLoS ONE 14(3): e0214391. https://doi.org/10.1371/journal.pone.0214391

Editor: Seyedali Mirjalili, Griffith University, AUSTRALIA

Received: October 10, 2018

Accepted: March 12, 2019

Published: March 26, 2019

Data Availability Statement: The summary statistics are available in the two articles: Clarke G, Eubanks D, Reid CK, et al. Overcoming Depression on the Internet (ODIN)(2): a randomized trial of a self-help depression skills program with reminders. Journal of Medical Internet Research 2005; 7: e16.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The omnibus test is commonly applied to evaluate the overall disparity between group means in ANOVA. Alternatively, linear contrasts are more informative in detecting specific pattern of mean differences that cannot be obtained via the omnibus test. This article concerns power and sample size calculations for contrast analysis with heterogeneous variances and budget concerns. Optimal allocation procedures for the Welch-Satterthwaite tests of standardized and unstandardized contrasts are presented to minimize the total sample size with the designated ratios, to meet a desirable power level for the least cost, and to attain the maximum power performance under a fixed cost. Currently available methods rely exclusively on simple allocation formula and direct rounding rule. The proposed allocation strategies combine the computing techniques of nonlinear optimization search and iterative screening process. Numerical assessments of a randomized control trial for the overcoming depression on the Internet are conducted to demonstrate and confirm that the approximate procedures do not guarantee optimal solution. The suggested approaches extend and outperform the existing findings in methodological soundness and overall performance. The corresponding computer algorithms are developed to implement the recommended power and sample size calculations for optimal contrast analysis.

Introduction

Within the context of analysis of variance (ANOVA), the omnibus F test is widely used for detecting the overall mean differences. Alternatively, many important research questions may be formulated as a linear combination of the population group means. Hence, a t test of individual contrast provides much more information than an omnibus hypothesis in assessing particular relation between the mean effects. Comprehensive exposition and further information can be found in Kutner et al. [1] and Maxwell and Delaney [2]. However, it has been noted in many actual applications that the homogeneous variances assumption of ANOVA is frequently violated. For example, Grissom [3], Rosopa, Schaffer, and Schroeder [4], and Ruscio and Roche [5] stressed that variances can be extremely different across treatment groups in clinical and psychological study. To account for the impact of variance heterogeneity, the Welch–Satterthwaite procedure of Satterthwaite [6] and Welch [7] is commonly recommended as an

alternative to the usual *t* test for detecting the substantive significance of a linear contrast. Accordingly, the contrast analysis under heterogeneity of variance is a generalization of the well-known Behrens–Fisher problem of testing the difference between two population means with unequal variances.

The general guidelines of experimental design and statistical analysis suggest that even the renowned test procedures do not warrant correct detection of treatment differences that are strongly expected or theoretically supported (Ioannidis [8], Moher, Dulberg, & Wells [9]). To prevent mistakenly dismissing an important contrast effect and to provide profound implications for ANOVA research, the underlying issues of power and sample size calculations must also be considered. It is prudent to emphasize that the traditional power and sample size procedures do not consider the cost implications. Notably, Allison et al. [10] advocated designing statistically powerful studies while minimizing costs. On the other hand, Marcoulides [11] emphasized the notion of maximizing power in designing studies under budget constraints. Although power and sample size procedures are available for Welch's test [12] of the difference between two means, Luh and Guo [13] noted that cost issues have not been incorporated in the Welch–Satterthwaite test for contrast analysis. Accordingly, Luh and Guo [13] described a formula for efficient sample size allocation in two scenarios. The first scenario is attaining a designated power with the minimum total cost, and the second scenario is maximizing the statistical power for a designated total cost. The suggested optimal sample sizes have a ratio that is proportional to the product of the ratio of contrast coefficients and the ratio of standard deviations divided by the square root of the ratio of unit sampling costs. The particular method is a direct extension of the optimal sample size formula in Dette and Munk [14] and Pentico [15] for detecting the difference between two means under the normality assumption.

A standard normal distribution can be viewed as a *t* distribution with an infinite number of degrees of freedom. Despite this large-sample argument, the importance of a Student's *t* distribution is well recognized in statistical applications, especially when the sample size is small. Note that the underlying notion of incorporating the cost concerns is because the time, money, and other resources are limited in all practical studies. The power and sample procedures [16–20] stress the theoretical principles of the approximate degree of freedom tests using estimated degrees of freedom. Specifically, power and sample size calculations for the Welch's [21] omnibus test have been presented in Jan and Shieh [16] and Shieh and Jan [17, 18]. On the other hand, Shieh and Jan [19] considered the problem of power and sample size for the Welch–Satterthwaite test of linear contrasts, but they did not consider budget issues. The related results in Jan and Shieh [20] are restricted for designing 2×2 factorial studies while minimizing financial costs. Therefore, these optimal sample size methods did not cover all the cost schemes for contrast analysis. To our knowledge, there has been no other optimal cost and allocation investigations for the Welch–Satterthwaite test of contrast analysis except for Luh and Guo [13]. As a generalization of the results in [16–20], the present study focuses on optimal contrast analysis by implementing the distributional properties of the Welch–Satterthwaite *t* statistic in cost and allocation evaluations. Accordingly, in addition to being able to contribute to the methodological development and understanding of the approximate degrees of freedom test procedure, it also facilitates the pedagogical and numerical comparisons of the suggested approaches and the methods of Luh and Guo [13] for optimal sample size determinations.

In addition to the unstandardized contrasts, the effect size reporting and interpretation practices suggest that the standardized effect sizes are useful when comparing results from multiple studies using measurement instruments whose raw units are not directly comparable, such as Fritz, Morris, and Richler [22], Lakens [23], and Takeshima et al. [24]. Notably, standardized contrasts of treatment effects and corresponding effect sizes in ANOVA have been

investigated by, among others, Olejnik and Algina [25], Rosenthal, Rosnow, and Rubin [26], and Steiger [27]. The prescribed sample size studies of linear contrasts did not include the more involved situations of standardized contrasts. Hence, it is of theoretical importance to extend the power and sample size calculations for conducting hypothesis testing of standardized contrasts.

In view of the importance of methodological justification and computational support, this article aims to present a systematic and thorough discussion for the Welch-Satterthwaite tests of standardized and unstandardized contrasts. Optimal allocation approaches are presented to minimize the total sample size with the designated ratios, to meet a desirable power level for the least cost, and to attain the maximum power performance under a fixed cost. An Internet depression intervention example is employed to demonstrate the features of the suggested approaches. To facilitate the recommended procedures in planning research designs, computer algorithms are offered for optimal power and sample size calculations with the designated allocation and cost schemes.

Methods

Linear contrasts

Consider the one-way ANOVA model in which Y_{ij} denotes the jth value of the response variable from the ith treatment group and the observations are assumed to be independent and normally distributed:

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

where μ_i and σ_i^2 are unknown parameters, $i = 1, \ldots, G\ (\geq 2)$ and $j = 1, \ldots, N_i$. In addition to or instead of the omnibus test, questions regarding a particular pattern of group differences can be tested with a contrast of mean values.

A contrast is defined as a linear combination of mean parameters

$$\psi = \sum_{i=1}^{G} l_i \mu_i, \tag{2}$$

where l_i are the linear coefficients with $\sum_{i=1}^{G} l_i = 0$. With the model assumption defined in Eq 1, an unbiased estimator $\hat{\psi}$ for the contrast ψ is of the form

$$\hat{\psi} = \sum_{i=1}^{G} l_i \bar{Y}_i \tag{3}$$

where $\bar{Y}_i = \sum_{j=1}^{N_i} Y_{ij}/N_i$ is the ith group sample mean and is an unbiased estimator of μ_i for $i = 1, \ldots, G$. Moreover, the contrast estimator $\hat{\psi}$ given in Eq 3 has the distribution

$$\hat{\psi} \sim N(\psi, \omega^2), \tag{4}$$

where $\omega^2 = Var(\hat{\psi}) = \sum_{i=1}^{G} l_i^2 \sigma_i^2/N_i$. An unbiased estimator $\hat{\omega}^2$ of ω^2 can be readily obtained by replacing the variance σ_i^2 in ω^2 with its unbiased estimator S_i^2:

$$\hat{\omega}^2 = \sum_{i=1}^{G} \frac{l_i^2 S_i^2}{N_i}, \tag{5}$$

where $S_i^2 = \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2/(N_i - 1)$ is the sample variance for $i = 1, \ldots, G$.

Test for difference. To appraise a linear contrast of the mean effects in terms of the hypothesis

$$H_0 : \psi = \psi_0 \text{ versus } H_1 : \psi \neq \psi_0, \tag{6}$$

the test statistic is of the form

$$T = \frac{\hat{\psi} - \psi_0}{\hat{\omega}}, \tag{7}$$

where ψ_0 is a constant. The Welch–Satterthwaite procedure suggests that under the null hypothesis H_0: $\psi = \psi_0$, the quantity T has a convenient approximate distribution

$$T \dot{\sim} t(v), \tag{8}$$

where $v = \{\sum_{i=1}^{G} l_i^2\sigma_i^2/N_i\}^2/\{\sum_{i=1}^{G} l_i^4\sigma_i^4/[N_i^2(N_i-1)]\}$ and $t(v)$ is a t distribution with degrees of freedom v. For inferential purposes, the degrees of freedom v is replaced by its counterpart $\hat{v}$ with direct substitution of $\{S_1^2, \ldots, S_G^2\}$ for $\{\sigma_1^2, \ldots, \sigma_G^2\}$ in v, where

$$\hat{v} = \frac{\{\sum_{i=1}^{G} l_i^2 S_i^2/N_i\}^2}{\sum_{i=1}^{G} l_i^4 S_i^4/[N_i^2(N_i-1)]} \tag{9}$$

The test rejects H_0 at the significance level α if $|T| > t_{1-\alpha/2}(\hat{v})$ where $t_{1-\alpha/2}(\hat{v})$ is the upper $100(\alpha/2)$ percentile of the t distribution $t(\hat{v})$.

Moreover, with the same theoretical arguments and analytic derivations, it can be shown that the statistic T has the general approximate distribution

$$T \dot{\sim} t(v, \Delta), \tag{10}$$

where $t(v, \Delta)$ is a noncentral t distribution with degrees of freedom v and noncentrality parameter

$$\Delta = \frac{\psi - \psi_0}{\omega}. \tag{11}$$

Also, the power function of the Welch–Satterthwaite test can be approximated by

$$\pi(\Delta) = P\{|t(v, \Delta)| > t_{1-\alpha/2}(v)\}. \tag{12}$$

Test for noninferiority and superiority. In addition to the two-sided test of difference for a contrast, it is of clinical importance to test the hypotheses for noninferiority and superiority between mean effects (Laster & Johnson [28], Mulla et al. [29], Piaggio et al. [30], Scott [31]). The problem of testing noninferiority and superiority can be unified by the following hypotheses when larger values of ψ are better:

$$H_0 : \psi \leq \psi_0 \text{ versus } H_1 : \psi > \psi_0 \tag{13}$$

where ψ_0 is the non-inferiority or superiority threshold (Fleming et al. [32], Gayet-Ageron et al. [33], Gayet-Ageron et al. [34], Gladstone & Vach [35], Wien [36]). When $\psi_0 < 0$, the rejection of the null hypothesis implies noninferiority against the reference margin. Whereas the rejection of the null hypothesis indicates superiority over the reference bound for $\psi_0 > 0$. The upper one-sided test procedure rejects the null hypothesis at the significance level α if

$T > t_{1-\alpha}(\hat{v})$ and the associated power function is expressed as

$$\pi(\Delta) = P\{|t(v, \Delta)| > t_{1-\alpha}(v)\}. \quad (14)$$

Related points to consider on switching between superiority and non-inferiority can be found in the report of the Committee for Proprietary Medicinal Products [37], Ganju and Rom [38], Lewis [39], and Murray [40].

Standardized contrasts

The usual linear contrast has advantages in understanding the meaning of effect size because the scale is the same as the original units of analysis. Alternatively, the standardized contrasts provide a natural interpretation of net effect that is critical to transform the magnitude of a treatment combination with respect to the metric of a response variable. A standardized contrast effect ψ^* is defined as

$$\psi^* = \frac{\psi}{\omega^*}, \quad (15)$$

where $\omega^{*2} = \sum_{i=1}^{G} l_i^2 \sigma_i^2 / q_i$ and $q_i = N_i/N_T$ for $i = 1, \ldots, G$, and $N_T = \sum_{i=1}^{G} N_i$. To detect a standardized contrast effect, a slightly different statistic than T is considered:

$$T^* = \frac{\hat{\psi}}{\hat{\omega}}. \quad (16)$$

Also, T^* has the general distribution

$$T^* \dot{\sim} t(v, \Delta^*), \quad (17)$$

where $\Delta^* = N_T^{1/2}\psi^*$.

Test for difference. For assessing the standardized contrast effects in terms of the hypothesis

$$H_0 : \psi^* = \psi_0^* \, versus \, H_1 : \psi^* \neq \psi_{0,}^* \quad (18)$$

the test statistic T^* has the distribution

$$T^* \dot{\sim} t(v, \Delta_0^*), \quad (19)$$

where $\Delta_0^* = N_T^{1/2}\psi_0^*$ and ψ_0^* is a constant. The null hypothesis is rejected at the significance level α if $T^* < t_{\alpha/2}(\hat{v}, \Delta_0^*)$ or $T^* > t_{1-\alpha/2}(\hat{v}, \Delta_0^*)$ where $t_{\alpha/2}(\hat{v}, \Delta_0^*)$ and $t_{1-\alpha/2}(\hat{v}, \Delta_0^*)$ are the lower and upper $100(\alpha/2)$ percentiles of the noncentral t distribution $t(\hat{v}, \Delta_0^*)$, respectively. The corresponding power function is

$$\pi^*(\Delta^*) = P\{t(v, \Delta^*) < t_{\alpha/2}(v, \Delta_0^*)\} + P\{t(v, \Delta^*) > t_{1-\alpha/2}(v, \Delta_0^*)\}. \quad (20)$$

Test for noninferiority and superiority. To perform the upper one-sided test for noninferiority and superiority in terms of

$$H_0 : \psi^* \leq \psi_0^* \, versus \, H_1 : \psi^* > \psi_0^*, \quad (21)$$

the test procedure rejects H_0 at the significance level α if $T^* > t_{1-\alpha}(\hat{v}, \Delta_0^*)$. Accordingly, the

power function is defined as

$$\pi^*(\Delta^*) = P\{t(v, \Delta^*) > t_{1-\alpha}(v, \Delta_0^*)\}. \quad (22)$$

The reference values ψ_0^* need to be prudently selected to reflect the planned tests of noninferiority or superiority with appropriate magnitude and sign.

Sample size calculations

During the planning stage of a research study, a question of essential interest is how many subjects are needed in order to have the desired power for conducting a scientifically meaningful analysis. To extend the applicability of contrast analysis, optimal sample size procedures are presented with respect to distinct allocation and cost concerns.

Sample size ratios are fixed. For advance planning of unstandardized and standardized contrast analysis, the prescribed power functions $\pi(\Delta)$ and $\pi^*(\Delta^*)$ can be employed to calculate the sample sizes $\{N_i, i = 1, \ldots, G\}$ needed to attain the specified power $1 - \beta$ for the chosen significance level α, null values ψ_0 and ψ_0^*, contrast coefficients $\{l_i, i = 1, \ldots, G\}$, and parameter values $\{(\mu_i, \sigma_i^2), i = 1, \ldots, G\}$. However, it is prudent to consider the design structure with a priori chosen sample size ratios $\{r_1, \ldots, r_G\}$ where $r_j = N_j/N_g \geq 1$, $j = 1, \ldots, G$, with the gth group has the smallest sample size N_g. Note that the sample size calculations in Shieh and Jan [19] are only applicable to the tests of difference for conventional linear contrasts. Hence, they did not consider the tests for the standardized contrasts.

The cost and effort to treat a subject often vary with treatment groups and it is sensible for researchers to take into account budget and resource constraints in research design. The total cost of an ANOVA study can be represented by the overhead cost and sampling costs through the following simple cost function

$$C_T = c_0 + \sum_{i=1}^{G} c_i N_i, \quad (23)$$

where c_O is the fixed overhead cost associated with the study, and c_i reflects unit sampling cost of each subject in group i for $i = 1, \ldots G$. Apparently, the cost assessment reduces to the evaluation of total number of subjects $C_T = N_T = \sum_{i=1}^{G} N_i$ when $c_O = 0$ and $c_i = 1$ for $i = 1, \ldots, G$. Under cost and power considerations, the following two scenarios arise naturally in choosing the optimal sample sizes.

Target power is fixed and total cost needs to be minimized. Despite the simple linear form of the objective cost function, the optimization process involves the designated power function as a nonlinear constraint. Thus, a closed form solution rarely exists for most situations. With the specifications of the significance level α, the desired power level $1 - \beta$, the null effect size, contrast coefficients, and the model parameters of group means and variance components, the suggested approach is composed of two key steps.

First, the preliminary set of sample sizes $\{N_{Pi}, i = 1, \ldots, G\}$ for attaining the desired power performance while minimizing the total cost can be obtained with the NLPQN subroutine of the SAS/IML [41] package. However, the sample sizes are treated as continuous variables in the optimization process. The resulting sample sizes are most likely non-integer values. In view of the discrete nature of sample sizes, a systematic evaluation is conducted to find the proper result in the second step. The screening process of Shieh and Jan [18] is extended for a wider range of sample size combinations. Specifically, power calculations and cost assessments are performed for a total of 4^G sample size sets $\{N_i, i = 1, \ldots, G\}$ with $N_i = [N_{P1}] - 1$, $[N_{P1}]$, $[N_{P1}] + 1$, or $[N_{P1}] + 2$ for $i = 1, \ldots G$, and $[M]$ denotes the integer part of M. Then, the optimal allocation $\{N_i^*, i = 1, \ldots, G\}$ is found through an inspection of the sample size combinations

that attain the desired power while giving the least cost. If more than one set yields the same amount of least cost, the one giving the largest power is reported.

In contrast to the proposed thorough search, Luh and Guo [13] showed that the potential optimal sample size ratio for contrast analysis of unstandardized means is proportional to the product of the ratio of contrast coefficients and the ratio of standard deviations divided by the square root of the ratio of unit sampling costs:

$$\gamma_i = \frac{N_i}{N_1} = \frac{|l_i|\sigma_i c_1^{1/2}}{|l_1|\sigma_1 c_i^{1/2}},\ i = 1,\ \ldots,\ G. \tag{24}$$

Note that the allocation ratios are derived with the standard normal Z statistic for known variances, rather than the t statistic under the assumption of unknown variances.

Total cost is fixed and actual power needs to be maximized. In addition to the prescribed design scheme, a problem of practical interest is to decide the best design in power performance when the total cost is fixed. Similar to the previous approach for optimal design, a two-step search procedure is performed. In this case, the specialized SAS/IML [41] NLPNRA subroutine is used to find the initial sample sizes $\{N_{P1}, \ldots, N_{PG}\}$ for the maximization of a nonlinear power function with the linear inequality cost constraint. The optimization algorithm assumes the sample sizes are continuous measurements and the computed outcomes $\{N_{P1}, \ldots, N_{PG}\}$ are extremely probable not integer values. To give the correct optimal solution, power and cost appraisals are performed in the second step for a total of 4^G sample size combinations $\{N_1, \ldots, N_G\}$ with $N_i = [N_{P1}] - 1$, $[N_{P1}]$, $[N_{P1}] + 1$, or $[N_{P1}] + 2$ for $i = 1, \ldots G$. Accordingly, the optimal allocation $\{N_1^*, \ldots, N_G^*\}$ is obtained through a detailed comparison of the sample size configurations that yields the greatest power while maintaining the restricted budget.

In this case, Luh and Guo [13] suggested the optimal sample size combination still has the allocation ratios given in Eq 24. The sample size of the first group is determined by $N_1 = (C_T - c_0)/(\sum_{i=1}^{G} c_i\gamma_i)$ and the other sample sizes are then computed with $N_i = N_1\gamma_i$ for $i = 2, \ldots, G$. It is unlikely that the sample sizes computed from the allocation ratios are whole numbers. The computed sample sizes need to be rounded up or down to the nearest integer and the outcomes are reported as the optimal sample sizes.

Results

To explicate the usefulness of the recommended exact approaches and associated computer programs, the overcoming depression on the Internet (ODIN) study of Clarke et al. [42] is exemplified for power and sample size calculations. This research was a three-arm randomized control trial with a usual treatment control group and two ODIN intervention groups receiving reminders through postcards or brief telephone calls.

For demonstration, Luh and Guo [13] suggested a specific comparison of the two intervention programs to the usual treatment without access to ODIN with respect to the mental component summary scores at 16-week. The hypothesis testing is formulated as H_0: $\psi \leq -4.2$ versus H_1: $\psi > -4.2$ with the linear coefficients $\{l_1, l_2, l_3\} = \{0.5, 0.5, -1\}$. For the three study conditions of mail reminder, telephone reminder, and control group, $\{N_1, N_2, N_3\} = \{75, 80, 100\}$, $\{\bar{Y}_1,\ \bar{Y}_2,\ \bar{Y}_3\} = \{34.7, 32.3, 35.5\}$, and $\{S_1^2,\ S_2^2,\ S_3^2\} = \{79.21, 57.76, 77.44\}$. It is shown that the contrast effect, estimated variance, and approximate degrees of freedom are $\hat{\psi} = -2$, $\hat{\omega}^2 = 1.2189$, and $\hat{v} = 200.4582$, respectively. Moreover, the observed test statistic $T = 1.9927$ and the p-value = 0.0238. Hence, the test concludes that the contrast effect is significantly larger than –4.2 at $\alpha = 0.05$.

For the purposes of power analysis and sample size determination, the abovementioned findings are employed to provide planning values of the model parameters and design characteristics for upcoming Internet depression intervention study. With these parameter settings, contrast coefficients {0.5, 0.5, −1}, sample sizes {75, 80, 100}, and $\psi_0 = -4.2$, the accompanying program shows that attained powers for the two-sided test and one-side test given in Eqs 6 and 13 are 0.5093 and 0.6335, respectively. The resulting powers are far less than the fairly common level of 0.80. Numerical computations reveal that the balanced group sample sizes of 183 and 144 are required to achieve the target power of 0.8 for the two-sided and one-side tests, respectively.

According to the cost-effectiveness study of Hollinghurst et al. [43], Luh and Guo [13] assumed that the fixed overhead cost $c_O = 0$ and the average operation costs $\{c_1, c_2, c_3\} = \{20, 50, 100\}$ as the unit sampling costs of the three treatment groups for future depression study. For the prescribed test for noninferiority, the approximate method of Luh and Guo [13] reported that the sample sizes {173, 93, 153} are required to attain the power performance of 0.80 with the least cost. Therefore, the total sample size and total cost are $N_T = 419$ and $C_T = 23{,}410$, respectively. It is also of fundamental interest to consider the optimal design problem in which the total number of subjects needs to be minimized. Luh and Guo [13] showed that the minimum sample sizes to assure the same power level of 0.80 are {98, 84, 193} with $N_T = 375$ and $C_T = 25{,}460$. Alternatively, the proposed approach suggests that the optimal sample sizes {173, 93, 152} and {97, 83, 193} are required to attain the designated power 0.80 with the least total cost and the smallest total sample size, respectively. The total sample size and total cost are $N_T = 418$ and $C_T = 23{,}310$ under the cost minimization consideration, whereas the corresponding results are $N_T = 373$ and $C_T = 25{,}390$ when minimum total sample size is desirable. The attained powers for the two sample size settings are 0.8000 and 0.8003, respectively, and they are nearly identical to the nominal level 0.80. For these two cases, Luh and Guo's [13] method consistently gave greater total costs and larger total sample sizes than the suggested algorithm.

For the scenario of finding the optimal allocation to maximize power performance when the total cost is fixed as 22,000, the sample sizes computed by Luh and Guo [13] are {162.43, 87.73, 143.65}. They suggested finding the appropriate sample sizes by rounding up or down to the nearest integer. Accordingly, their chosen sample sizes are {162, 87, 144} with the total cost $C_T = 21{,}990$. When a computer is not available, the checking procedure entails laborious and tedious calculations especially for four or more groups. Instead, the optimal sample size allocation computed by the proposed approach is {160, 88, 144} which perfectly meets the planned budget. Moreover, exact computation shows that the resulting power 0.7795 of the optimal structure is larger than the power 0.7793 attained by the prescribed sample sizes {162, 87, 144}. Hence, the proposed algorithm is superior to the approximate procedure of Luh and Guo [13]. Generally, the computations of optimal solutions can be simplified by the approximate methods without much loss in accuracy, especially when the sample sizes are large. However, the proposed approaches will produce more accurate results across all sample sizes.

To demonstrate the hypothesis testing, power computation, and sample size determination for the standardized contrasts, the comparison of mental component summary scores between the depression interventions is analyzed next. It follows from the definitions of the unstandardized contrast ψ and the standardized contrast ψ^* that a working value of ψ_0^* is $\psi_0/(N_T\hat{\omega}^2)^{1/2} = -0.2382 \doteq -0.25$. For simplicity's sake, the null standardized effect is set as $\psi_0^* = -0.25$. Then, the hypothesis testing in terms of the standardized measure is formulated by H_0: $\psi^* \leq -0.25$ versus H_1: $\psi^* > -0.25$. With the given data, the computations show that the standardized test statistic $T^* = -1.8115$, the critical value $t_{0.95}(200.4582, -3.9922) = -2.3390$,

and the p-value = 0.0149. It is concluded that the standardized effect is significantly larger than –0.25 at $\alpha = 0.05$.

With the parameter settings $\{\mu_1, \mu_2, \mu_3\} = \{34.7, 32.3, 35.5\}$ and $\{\sigma_1^2, \sigma_2^2, \sigma_3^2\} = \{79.21, 57.76, 77.44\}$, the statistical power associated with the previous sample size combination {75, 80, 100} is 0.6988. For a balanced structure, it can be shown that the minimum sample size set {94, 94, 94} is necessary to attain the designated power of 0.8. In this case, the total sample size is $N_T = 282$ and the total cost is $C_T = 15{,}980$ for the fixed overhead cost $c_O = 0$ and the average operation costs $\{c_1, c_2, c_3\} = \{20, 50, 100\}$. To attain the designated power 0.80 with the minimum total cost, the suggested procedure yields the optimal sample size scheme {305, 7, 15} with the total sample size $N_T = 327$ and total cost $C_T = 7{,}950$. On the other hand, the suggested allocation {73, 62, 143} incurs the least total sample size $N_T = 278$ with $C_T = 18{,}860$. In this case, the balanced design is not the optimal solution for either consideration of minimum total sample size or minimum total cost. When the maximum total cost is 22,000, the proposed optimal sample size structure is {815, 22, 46} with $N_T = 883$ and $C_T = 22{,}000$.

Note that all the numerical results of the optimal power and sample size procedures were computed with the supplemental SAS/IML algorithms. For ease of application, two different sets of computer programs are presented for the standardized and unstandardized contrast analysis.

Conclusions and discussion

The Welch–Satterthwaite statistic and the associated approximate t distribution have an important utility in accommodating the impact of heterogeneity of variance in statistical inference. The technical account of diverse hypothesis-testing frameworks enhances the theoretical implication and practical usefulness of the Welch–Satterthwaite test for contrast analysis in the detection of difference or inferiority/superiority. Moreover, the integrated document of different contrast effect sizes facilitates the reporting and interpretation of important finding in standardized measure scaled by the associated variabilities and design characteristics or in simple magnitude expressed in the same metric as the original units of analysis. One important implication of this research is that the essence of the Welch–Satterthwaite procedure is properly recognized in related power and sample size calculations without a normal simplification. Nonlinear optimization routines and systematic numerical evaluations are synthesized to give optimal sample size allocations for contrast analysis. According to the analytic examination and numerical assessment, the suggested procedures ultimately outperform the existing sample size methods based on the normal approximation and integer rounding. Essentially, the collection of computer programs covers both the two-sided and one-sided hypothesis testing for the two distinct formulations of standardized and unstandardized contrasts. The presented appraisals of statistical power, sample size, and financial budget should be useful for researchers to justify their allocation strategy and project support in planning research design.

The general formulation of a linear contrast of group means permits a wide range of research hypotheses to be tested in ANOVA. To enhance the usefulness of contrast analysis under heterogeneity of variance, this article addresses the problem of optimal sample size calculations for the Welch–Satterthwaite test with cost constraints. The present study has three essential features. First, the two-sided and one-sided test procedures are presented for both the standardized and unstandardized contrasts in ANOVA under the heterogeneous variances assumption. Second, optimal sample size approaches are proposed for the two essential problems that when the target power is fixed and total cost needs to be minimized and when the total cost is fixed and actual power needs to be maximized. Third, computer codes are

presented to implement the power and sample size computations of the Welch–Satterthwaite procedures. In sum, this study contributes to the current literature for optimal research designs by alleviating the limitations of existing investigations and extending the usefulness of contrast analysis in ANOVA under variance heterogeneity.

Supporting information

S1 File. SAS/IML programs for performing the tests of linear contrast.
(PDF)

S2 File. SAS/IML programs for performing the tests of standardized contrast.
(PDF)

Author Contributions

Conceptualization: Show-Li Jan, Gwowen Shieh.

Formal analysis: Show-Li Jan, Gwowen Shieh.

Investigation: Show-Li Jan, Gwowen Shieh.

Methodology: Show-Li Jan, Gwowen Shieh.

Project administration: Show-Li Jan.

Software: Gwowen Shieh.

Writing – original draft: Gwowen Shieh.

Writing – review & editing: Show-Li Jan, Gwowen Shieh.

References

1. Kutner M. H., Nachtsheim C. J., Neter J., & Li W. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw Hill.
2. Maxwell S. E., & Delaney H. D. (2004). *Designing experiments and analyzing data*: *A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
3. Grissom R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68, 155–165. PMID: 10710850
4. Rosopa P. J., Schaffer M. M., & Schroeder A. N. (2013). Managing heteroscedasticity in general linear models. *Psychological Methods*, 18, 335–351. https://doi.org/10.1037/a0032553 PMID: 24015776
5. Ruscio J., & Roche B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology*, 8, 1–11.
6. Satterthwaite F. E. (1946). An approximate distribution of estimate of variance components. *Biometrics Bulletin*, 2, 110–114. PMID: 20287815
7. Welch B. L. (1947). The generalization of Students' problem when several different population variances are involved. *Biometrika*, 34, 28–35. PMID: 20287819
8. Ioannidis J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
9. Moher D., Dulberg C. S., & Wells G. A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association*, 272, 122–124. PMID: 8015121
10. Allison D. B., Allison R. L., Faith M. S., Paultre F., & Pi-Sunyer X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods*, 2, 20–33.
11. Marcoulides G. A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational Statistics*, 18, 197–206.
12. Welch B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350–362.

13. Luh W. M., & Guo J. H. (2016). Sample size planning for the non-inferiority or equivalence of a linear contrast with cost considerations. *Psychological Methods*, 21, 13–34. https://doi.org/10.1037/met0000039 PMID: 26121080

14. Dette H., & Munk A. (1997). Optimum allocation of treatments for Welch's test in equivalence assessment. *Biometrics*, 53, 1143–1150. PMID: 9333344

15. Pentico DW. On the determination and use of optimal sample sizes for estimating the difference in means. The American Statistician. 1981; 35: 41–42.

16. 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. https://doi.org/10.1111/bmsp.12006 PMID: 23316952

17. Shieh G., & Jan S. L. (2013). Determining sample size with a given range of mean effects in one-way heteroscedastic analysis of variance. *Journal of Experimental Education*, 81, 281–294.

18. Shieh G., & Jan S. L. (2015). Optimal sample size allocation for Welch's test in one-way heteroscedastic ANOVA. *Behavior Research Methods*, 47, 374–383. https://doi.org/10.3758/s13428-014-0477-8 PMID: 24903689

19. Shieh G., & Jan S. L. (2015). Power and sample size calculations for testing linear combinations of group means under variance heterogeneity with applications to meta and moderation analyses. *Psicologica*, 36, 367–390.

20. Jan S. L., & Shieh G. (2016). A systematic approach to designing statistically powerful heteroscedastic 2 × 2 factorial studies while minimizing financial costs. *BMC Medical Research Methodology*, 16, 114. https://doi.org/10.1186/s12874-016-0214-3 PMID: 27578357

21. Welch B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330–336.

22. 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.

23. Lakens D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863 PMID: 24324449

24. Takeshima N., Sozu T., Tajika A., Ogawa Y., Hayasaka Y., & Furukawa T. A. (2014). Which is more generalizable, powerful and interpretable in meta-analyses, mean difference or standardized mean difference?. *BMC Medical Research Methodology*, 14, 30. https://doi.org/10.1186/1471-2288-14-30 PMID: 24559167

25. Olejnik S., & Algina J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241–286. https://doi.org/10.1006/ceps.2000.1040 PMID: 10873373

26. Rosenthal R., Rosnow R. L., & Rubin D. B. (2000). *Contrasts and effect sizes in behavioral research*: *A correlational approach*. New York: Cambridge University Press.

27. Steiger J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164–182. https://doi.org/10.1037/1082-989X.9.2.164 PMID: 15137887

28. Laster L. L., & Johnson M. F. (2003). Non-inferiority trials: the 'at least as good as' criterion. *Statistics in Medicine*, 22, 187–200. PMID: 12520556

29. Mulla S. M., Scott I. A., Jackevicius C. A., You J. J., & Guyatt G. H. (2012). How to use a noninferiority trial: Users' guides to the medical literature. *Journal of the American Medical Association*, 308, 2605–2611. https://doi.org/10.1001/2012.jama.11235 PMID: 23268519

30. Piaggio G., Elbourne D. R., Altman D. G., Pocock S. J., Evans S. J., & Consort Group. (2006). Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. *Journal of the American Medical Association*, 295, 1152–1160. https://doi.org/10.1001/jama.295.10.1152 PMID: 16522836

31. Scott I. A. (2009). Non-inferiority trials: Determining whether alternative treatments are good enough. *Medical Journal of Australia*, 190, 326–330. PMID: 19296815

32. Fleming T. R., Odem-Davis K., Rothmann M. D., & Li Shen Y. (2011). Some essential considerations in the design and conduct of non-inferiority trials. *Clinical Trials*, 8, 432–439. https://doi.org/10.1177/1740774511410994 PMID: 21835862

33. Gayet-Ageron A., Agoritsas T., Rudaz S., Courvoisier D., & Perneger T. (2015). The choice of the non-inferiority margin in clinical trials was driven by baseline risk, type of primary outcome, and benefits of new treatment. *Journal of Clinical Epidemiology*, 68, 1144–1151. https://doi.org/10.1016/j.jclinepi.2015.01.017 PMID: 25716902

34. Gayet-Ageron A., Jannot A. S., Agoritsas T., Rudaz S., Combescure C., & Perneger T. (2016). How do researchers determine the difference to be detected in superiority trials? Results of a survey from a panel of researchers. *BMC Medical Research Methodology*, 16, 89. https://doi.org/10.1186/s12874-016-0195-2 PMID: 27473336

35. Gladstone B. P., & Vach W. (2014). Choice of non-inferiority (NI) margins does not protect against degradation of treatment effects on an average–an observational study of registered and published NI trials. *PLoS ONE*, 9, e103616. https://doi.org/10.1371/journal.pone.0103616 PMID: 25080093

36. Wiens B. L. (2002). Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled Clinical Trials*, 23, 2–14. PMID: 11852160

37. Committee for Proprietary Medicinal Products (2001). Points to consider on switching between superiority and non-inferiority. *British Journal of Clinical Pharmacology*, 52, 223–228. https://doi.org/10.1046/j.0306-5251.2001.01397-3.x PMID: 11560553

38. Ganju J., & Rom D. (2017). Non-inferiority versus superiority drug claims: the (not so) subtle distinction. *Trials*, 18, 278. https://doi.org/10.1186/s13063-017-2024-2 PMID: 28619049

39. Lewis J. A. (2001). Switching between superiority and non-inferiority: an introductory note. *British Journal of Clinical Pharmacology*, 52, 221–221. PMID: 11560552

40. Murray G. D. (2001). Switching between superiority and non-inferiority. *British Journal of Clinical Pharmacology*, 52, 219–219. https://doi.org/10.1046/j.0306-5251.2001.01397.x PMID: 11560551

41. Institute SAS (2014). *SAS/IML User's Guide*, *Version 9.3*. Cary, NC: SAS Institute Inc.

42. Clarke G., Eubanks D., Reid C. K., O'Connor E., DeBar L. L., Lynch F., . . . & Gullion C. (2005). Overcoming Depression on the Internet (ODIN)(2): a randomized trial of a self-help depression skills program with reminders. *Journal of Medical Internet Research*, 7, e16. https://doi.org/10.2196/jmir.7.2.e16 PMID: 15998607

43. Hollinghurst S., Peters T. J., Kaur S., Wiles N., Lewis G., & Kessler D. (2010). Cost-effectiveness of therapist-delivered online cognitive–behavioural therapy for depression: Randomised controlled trial. *The British Journal of Psychiatry*, 197, 297–304. https://doi.org/10.1192/bjp.bp.109.073080 PMID: 20884953