Behav Res (2011) 43:1014–1022
DOI 10.3758/s13428-011-0095-7

Optimal sample sizes for Welch's test under various allocation and cost considerations

Show-Li Jan · Gwowen Shieh

Published online: 22 April 2011

Abstract The issue of the sample size necessary to ensure adequate statistical power has been the focus of considerableattention in scientific research. Conventional presentations of sample size determination do not consider budgetary and participant allocation scheme constraints, although there is some discussion in the literature. The introduction of additional allocation and cost concerns complicates study design, although the resulting procedure permits a practical treatment of sample size planning. This article presents exact techniques for optimizing sample size determinations in the context of Welch (*Biometrika, 29,* 350–362, 1938) test of the difference between two means under various design and cost considerations. The allocation schemes include cases in which (1) the ratio of group sizes is given and (2) one sample size is specified. The cost implications suggest optimally assigning subjects (1) to attain maximum power performance for a fixed cost and (2) to meet adesignated power level for the least cost. The proposed methods provide useful alternatives to the conventional procedures and can be readily implemented with the developed R and SAS programs that are available as supplemental materials from brm.psychonomic-journals.org/content/supplemental.

Keywords Behrens–Fisher problem · Power · Study design

Electronic supplementary material The online version of this article (doi:10.3758/s13428-011-0095-7) contains supplementary material, which is available to authorized users.

S.-L. Jan
Department of Applied Mathematics,
Chung Yuan Christian University,
Chungli, Taiwan 32023, Republic of China
e-mail: sljan@math.cycu.edu.tw

G. Shieh (✉)
Department of Management Science,
National Chiao Tung University,
1001 Ta Hsueh Road,
Hsinchu, Taiwan 30050, Republic of China
e-mail: gwshieh@mail.nctu.edu.tw

The a priori determination of a proper sample size necessary to achieve some specified power is an important problem frequently encountered in practical studies. To make inferences about differences between two normal population means, the hypothesis-testing procedure and corresponding sample size formula are well known and easy to apply. For important guidance, see the comprehensive treatments in Cohen (1988) and Murphy and Myors (2004). In the statistical literature, comparison of the means of two normal populations with unknown and possibly unequal variances has been the subject of much discussion and is well recognized as the Behrens–Fisher problem (Kim & Cohen, 1998). The existence and importance of violation of the assumption of homogeneity of variance in clinical research settings are also addressed in Grissom (2000). The practical importance and methodological complexity of the problem have occasioned numerous attempts to develop various procedures and algorithms for resolving the issue. Notably, several studies have shown that Welch's (1938) approximate degrees of freedom approach offers a reasonably accurate solution to the Behrens–Fisher problem. Therefore, Welch's procedure is routinely introduced in elementary statistics courses and textbooks. Moreover, some popular statistical computer packages, such as SAS and SPSS, have implemented the method for quite some time. In practice, power analyses and sample size calculations are often critical for investigators to credibly address specific research hypotheses and confirm differences. Thus, the planning of sample size should be included as an

integral part in the study design. Accordingly, it is of practical interest and fundamentalimportance to be able to perform these tasks in the context of the Behrens–Fisher problem. The essential question is how to determine sample sizes optimally under different allocation and cost considerations that call for independent random samples from two normal populations with possibly unequal variances.

Conventional studies of power and sample size have not addressedmatters of allocationrestriction and cost efficiency, although researchers have been exploring design strategies that take into account the impact of different constraints of the sample scheme and project funding while maintaining adequate power. Specifically, the allocation ratio of group sizes was fixed in the calculation of sample size for comparing independent proportions in Fleiss, Tytun, and Ury (1980), while Heilbrun and McGee (1985) considered sample size determination for the comparison of normal means with a known ratio of variances and one sample size being specified in advance.In an actual experiment, however, the available resources are generally limited, and it may require different amounts of effort and costs to recruit subjects for the treatment and the control groups. Assuming homogeneous variances, Nam (1973) presented optimal sample sizes to maximize power for the comparison of the treatment and control under budget constraints. Conversely, Allison, Allison, Faith, Paultre, and Sunyer (1997) advocated designing statistically powerful studies while minimizing costs. Interested readers are referred to recent articles by Bacchetti, McCulloch, and Segal (2008) and Bacchetti (2010) for alternative viewpoints and related discussions.

Within the framework of the Behrens–Fisher problem, assuming a desired sample size ratio, Schouten (1999) derived an approximate formula for computing sufficient sample size for a selected power. In addition, in Schouten (1999), a simplified sample size formula was proposed to minimize the total cost when the cost of treating a subject varies with experimental groups. Also, Lee (1992) determined the optimal sample sizes for a designated power so that the total sample size is minimized. It is important to note that the setting in Lee can be viewed as a special case of Schouten. However, unlike the exact approach of Lee, the presentation of Schouten involved several approximations, including the use of a normal distribution, which does not conform to the notion of a t distribution with *approximate degrees of freedom* proposed in Welch (1938). Alternatively, Singer (2001) modified the simple formula of Schouten by replacing the percentiles of the standard normal distribution with those of a t distribution with approximate degrees of freedom. Unfortunately, the resulting formulation is questionable on account of its absence of theoretical justification. Detailed analytical and empirical examinations are presented later to demonstrate the underlying drawbacks associated with the approximate procedures of Schouten and Singer. Moreover, Luh and Guo (2007), Guo and Luh (2009), and Luh and Guo (2010) extended the approximations of Schouten and Singer to the two-sample trimmed mean test with unequal variances under allocation and cost considerations. Basically, when the trimming proportion is 0, the procedures of Guo and Luhare applicable for the Behrens–Fisher problem. However, their procedures are still approximate in nature and possess the same disadvantages of Schouten's and Singer's. More important, the algorithms employed by Guo and Luhfail to take into account the underlying metric of integer sample sizes and often lead to suboptimal results. From a methodological standpoint, the results in Schouten, Singer, Luh and Guo (2007), Guo and Luh, and Luh and Guo (2010) should be reexamined with technical clarifications and exact computations. Nonetheless, our calculations not only show that the prescribed approximate methods do not guarantee giving correct optimal sample sizes, but also reveal that some of the optimal sample sizes reported in the empirical illustrations of Lee are actually suboptimal. Due to the discrete character of sample size, it requires a detailed inspection of sample size combinations to find the optimal allocation that attains the desired power while giving the least total sample size. This extra step and resulting merit in sample size determination is not considered by Lee. The theoretical and numericalexaminations conducted here provide a comprehensive comparison of the various procedures available to date.In short, the accuracy of the existing sample size procedures for the Behrens–Fisher problem can be further improved by adapting an exact and refined approach.

As was described above, there are important and useful considerations or strategies for study design other than the minimization of total sample size or total cost. Since Welch's (1938) approach to the Behrens–Fisher problem is so entrenched, it is prudent to present a comprehensive exposition of design configurations in terms of diverse allocation schemes and budget constraints. Here, exact methods are presented to give proper sample sizes either when the ratio of group sizes is fixed in advance or when one sample size is fixed. In addition, detailed procedures are provided to determine the optimal sample sizes that maximize the power for a given total cost and that minimize the cost for a specified power. More important, the corresponding computer algorithms are developed to facilitate computation of the exact necessary sample sizes in actual applications.

Due to the prospective nature of advance research planning, it is difficult to assess the adequacy of selected configurations for model parameters in sample size calculations. The general guideline suggests that typical sources such as previously published research and successful pilot studies can offer plausible and reasonable planning values

for the vital model characteristics (Thabane et al., 2010). However, the potential deficiency of using a pilot sample variance to compute the sample size needed to achieve the planned power for one- and two-sample *t*-tests has been examined by, among others, Browne (1995) and Kieser and Wassmer (1996). They showed that the sample sizes provided by the traditional formulas are too small, since they neglect the imprecise nature of a variance estimate. Note that all standard sample size procedures share the same fundamental weakness when sample variance estimates are used for the underlying population parameters. However, the issue is more involved, and a detailed discussion of this topic is beyond the scope of the present study. The interested reader is referred to Browne, Kieser, and Wassmer, and the references therein for further details.

The Welch test

As part of a continuing effort to improve the quality of research findings, this research contributes to the derivation and evaluation for sample size methodology of Welch's (1938) approximate *t* test for the Behrens–Fisher problem. Consider independent random samples from two normal populations with the following formulations:

$$X_{ij} \sim N\left(\mu_i, \sigma_i^2\right),$$

where μ_1, μ_2, σ_1^2, and σ_2^2 are unknown parameters, $j = 1, \ldots, N_i$, and $i = 1$ and 2. For detecting the group effect in terms of the hypothesis H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, the well-known Welch's *t* statistic has the form

$$V = \frac{\overline{X}_1 - \overline{X}_2}{\left(S_1^2/N_1 + S_2^2/N_2\right)^{1/2}},$$

where $\overline{X}_1 = \sum_{j=1}^{N_1} X_{1j}/N_1, \overline{X}_2 = \sum_{j=1}^{N_2} X_{2j}/N_2, S_1^2 = \sum_{j=1}^{N_1} \left(X_{1j} - \overline{X}_1\right)^2/(N_1 - 1)$, and $S_2^2 = \sum_{j=1}^{N_2} \left(X_{2j} - \overline{X}_2\right)^2/(N_2 - 1)$. Under the null hypothesis H_0: $\mu_1 = \mu_2$, Welch (1938) proposed the approximate distribution for *V*:

$$V \dot{\sim} t(\widehat{\nu}), \tag{1}$$

where $t(\widehat{\nu})$ is the *t* distribution with degrees of freedom $\widehat{\nu}$ and $\widehat{\nu} = \widehat{\nu}\left(N_1, N_2, S_1^2, S_2^2\right)$, with

$$1/\widehat{\nu} = \frac{1}{N_1 - 1}\left\{\frac{S_1^2/N_1}{S_1^2/N_1 + S_2^2/N_2}\right\}^2 + \frac{1}{N_2 - 1}\left\{\frac{S_2^2/N_2}{S_1^2/N_1 + S_2^2/N_2}\right\}^2.$$

Hence, H_0 is rejected at the significance level α if $|V| > t_{\widehat{\nu},\alpha/2}$, where $t_{\widehat{\nu},a/2}$ is the upper $100(\alpha/2)$th percentile of the *t* distribution $t(\widehat{\nu})$. The same notion was independently suggested by Smith (1936) and Satterthwaite (1946), and the test is sometimes referred to as the Smith–Welch–Satterthwaite test.

It is important to emphasize that the degrees of freedom $\widehat{\nu}$ is bounded by the smaller of $N_1 - 1$ and $N_2 - 1$ at one end and by $N_1 + N_2 - 2$ at the other—that is, $\text{Min}(N_1 - 1, N_2 - 1) \leq \widehat{\nu} \leq N_1 + N_2 - 2$. Because the critical value $t_{df,\ \alpha/2}$ decreases as *df* increases, the approximate critical value $t_{\widehat{\nu},a/2}$ is slightly larger than that of the two-sample *t* test $t_{N_1+N_2-2,\alpha/2}$ under homogeneity of variance assumptions. Although the differences between the two critical value saresmall with moderate to large sample sizes, they reflect the conceptual distinction between the corresponding Welch's *t* test and the regular two-sample *t* test. Note that a standard normal distribution can be viewed as a *t* distribution with an infinite number of degrees of freedom. However, the close resemblance between a standard normal distribution and a *t* distribution never causesthe introductory courses or textbooks to omit the coverage of Student's *t* distribution. Therefore, the theoretical distinction and implication between the critical value $t_{N_1+N_2-2,a/2}$ and a standard normal critical value $z_{\alpha/2}$ is highlyanalogous to that between $t_{\widehat{\nu},a/2}$ and $t_{N_1+N_2-2,a/2}$. Ultimately, the *t* approximation with the approximate degrees of freedom given in Eq. 1 serves as the prime solution to the Behrens–Fisher problem.

Although the underlying normality assumption in the above-mentioned two-sample location problem provides a convenient and useful setup, the exact distribution of Welch's test statistic *V* is comparatively complicated and may be expressed in different forms (see Wang, 1971, Lee & Gurland, 1975, and Nel, van der Merwe, & Moser, 1990, for technical derivation and related details). For ease of presentation, we need to develop some notation. It follows from the fundamental assumption that $Z = \left(\overline{X}_1 - \overline{X}_2\right)/\sigma \sim N(\delta, 1)$, $\delta = \mu_d/\sigma$, $\mu_d = (\mu_1 - \mu_2)$, $\sigma^2 = \sigma_1^2/N_1 + \sigma_2^2/N_2$, $W = (N_1 - 1)S_1^2/\sigma_1^2 + (N_2 - 1)S_2^2/\sigma_2^2 \sim \chi^2(N_1 + N_2 - 2)$ and $B = \left\{(N_1 - 1)S_1^2/\sigma_1^2\right\}/W \sim \text{Beta}\{(N_1 - 1)/2, (N_2 - 1)/2\}$. Thus, we consider the following alternative expression of *V* for its ease of numerical investigation:

$$V = \frac{T}{H^{1/2}}, \tag{2}$$

where $T = Z/\{W/(N_1 + N_2 - 2)\}^{1/2} \sim t(N_1 + N_2 - 2, \delta)$, $t(N_1 + N_2 - 2, \delta)$ is the noncentral *t* distribution with degrees of freedom $N_1 + N_2 - 2$, and noncentrality parameter δ, and $H = [(\sigma_1^2/N_1)\{B/p\} + (\sigma_2^2/N_2)\{(1 - B)/(1 - p)\}]/\sigma^2$ with $p = (N_1 - 1)/(N_1 + N_2 - 2)$. Note that the random variables *Z*, *W*, and *B* are mutually independent. Hence, *T* and *B* are independent. Also, it is important to note that $1/\widehat{\nu} = B_1^2/(N_1 - 1) + B_2^2/(N_2 - 1)$ where $B_2 =$

$1-B_1$ and $B_1 = [(\sigma_1^2/N_1)\{B/p\}]/[(\sigma_1^2/N_1)\{B/p\} + (\sigma_2^2/N_2)\{(1-B)/(1-p)\}]$. Hence, both H and are functions of the random variable B.

With the prescribed distributional properties in Eq. 2, the associated power function of V is denoted by

$$\pi(\mu_d, \sigma_1^2, \sigma_2^2, N_1, N_2) = P\left\{|V| > t_{\hat{\nu}, \alpha/2}\right\}$$
$$= P\left\{|T| > t_{\hat{\nu}, \alpha/2} \cdot H^{1/2}\right\} \quad (3)$$

The numerical computation of exact power requires the evaluation of the cumulative distribution function of a noncentral t variable and the one-dimensional integration with respect to a beta probability distribution function. Since all related functions are readily embedded in major statistical packages, the exact computations can be conducted with current computing capabilities. To determine sample size, the power function can be employed to calculate the sample sizes (N_1, N_2) needed to attain the specified power 1-β for the chosen significance level α and parameter values $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. Clearly, the power function is rather complex, and it usually involves an iterative process to find the solution, because both random variables V and $t_{\hat{\nu}, \alpha/2}$ are functions of the sample sizes (N_1, N_2). In order to enhance the applicability of sample size methodology and the fundamental usefulness of Welch's (1938) procedure, in subsequent sections this study considers design configurations allowing for different allocation constraints and cost considerations. The R(R Development Core Team, 2010) and SAS/IML (SAS Institute, 2008a) programs employed to perform the corresponding sample size calculations are available in the supplementary files.

Allocation constraints

Since there may be several possible choices of sample sizes N_1 and N_2 that satisfy the chosen power level in the process of sample size calculations, it is prudent to consider an appropriate design that permits unique and optimal result. The following two allocation constraints are considered because of their potential usefulness. First, the ratio $r = N_2/N_1$ between the two group sizes may be fixed in advance, so the task is to decidetheminimum sample size $N_1(N_2 = rN_1)$ required to achieve the specified power level. Second, one of the two sample sizes—say, N_2—may be pre-assigned, and so the smallest size N_1 required to satisfy the designated power should be found.

Sample size ratio is fixed

Assume that the sample size ratio $r = N_2/N_1$ is fixed in advance. To facilitate computation, without loss of generality, the ratio can be taken as $r \geq 1$. Then the power function $\pi(\mu_d, \sigma_1^2, \sigma_2^2, N_1, N_2)$ of V becomes a strictly monotone function of N_1 when all other factors are treated as constants. A simple incremental search can be conducted to find the minimum sample size N_1 needed to attain the specified power 1-β for the chosen significance level α and parameter values $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. To simplify the computation, the large-sample normal approximation $V \sim N(\delta, 1)$ can be used to provide initial values to start the iteration. Specifically, the starting sample size N_{1Z} computed by the normal approximation would be the smallest integer that satisfies the inequality

$$N_{1Z} \geq (\sigma_1^2 + \sigma_2^2/r)(z_{\alpha/2} + z_\beta)^2/\mu_d^2, \quad (4)$$

where $z_{\alpha/2}$ and z_β are the upper $100(\alpha/2)$th and $100\cdot\beta$th percentiles of the standard normal distribution, respectively.

For illustration, when $\mu_d = 1$, $\alpha = .05$ and $1 - \beta = .90$, the sample sizes N_1 and $N_2 = r \cdot N_1$ are presented in Table 1 for selected values of $r = 1$, 2, and 3, $\sigma_1 = 1/3$, 1/2, 1, 2, and 3, and $\sigma_2 = 1$. The actual power is also listed, and the values are marginally larger than the nominal level .90. Note that SAS procedure PROC POWER (SAS Institute, 2008b) provides the same feature to find the optimal sample sizes N_1 and N_2 with a given sample size ratio. However, it does not accommodate the extended settings in which one of the sample sizes is fixed and the more involved cost concernsthat we consider next.

One sample size is fixed

For ease of exposition, the sample sizeN_2 of the second group is held constant. Just as in the previous case, the

Table 1 Computed sample sizes (N_1, N_2) and actual power when sample size ratio $r = N_2/N_1$ is fixed with $\mu_d = 1$, $\alpha = .05$ and $1 - \beta = .90$

σ_1:σ_2	1/3:1			1/2:1			1:1			2:1			3:1		
r	N_1	N_2	Power	N_1	N_2	Power	N_1	N_2	Power	N_1	N_2	Power	N_1	N_2	Power
1	14	14	.9137	15	15	.9088	23	23	.9121	54	54	.9007	107	107	.9009
2	8	16	.9300	9	18	.9131	17	34	.9033	49	98	.9009	102	204	.9012
3	6	18	.9379	7	21	.9075	16	48	.9143	48	144	.9048	100	300	.9004

minimum sample size N_1 needed to ensure the specified power $1-\beta$ can be found by a simple iterative search for the chosen significance level α and parameter values $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. In this case, the starting sample size N_{1Z}, based on the normal approximation, is the smallest integer that satisfies the inequality

$$N_{1Z} \geq \sigma_1^2/\{\mu_d^2/(z_{a/2} + z_\beta)^2 - \sigma_2^2/N_2\}. \quad (5)$$

However, it should be noted that this may be problematic when a small value of N_2 is chosen. If $N_2 < \sigma_2^2/\{\mu_d^2/(z_{\alpha/2} + z_\beta)^2\}$, then the initial value N_{1Z} is negative, which is obviously unrealistic. Moreover, for $N_2 \doteq \sigma_2^2/\left\{\mu_d^2/(z_{\alpha/2} + z_\beta)^2\right\}$, the resulting N_{1Z} and N_1values are unbounded, and the results do not have practical value. Accordingly, Table 2 presents the computed sample size N_1 and the actual power levels with the chosen valueN_2 for the same settings with μ_d = 1, α = .05, 1-β = .90, and the variance combinations in Table 1.

Cost considerations

With limited research funding, it is desirable to consider the cost and effectiveness issues during the planning stage. In addition, the costs of obtaining subjects of treatment and control groups are not necessarily the same. Suppose that c_1 and c_2 are the costs per subject in the first and second groups, respectively; then, the total cost of the experiment is $C = c_1N_1 + c_2N_2$. The following two questions arise with considerable frequency in sample size determinations. First, given a fixed amount of money, what is the maximum power that the design can achieve? Second, assuming a preferred degree of power, what is the design that costs the least? In both cases, equal sample sizes for the two groups do not necessarily yield the optimal solution (Allison et al., 1997). Consequently, optimally unbalanced designs are more efficient, and a detailed and systematic approach to sample size allocation is required.

With the simplified asymptotic approximation of Welch's test $V \dot{\sim} N(\delta, 1)$, the optimal allocation isobtained-for the prescribed two scenarios when the ratio of the sample sizes assumes the equality

$$\frac{N_2}{N_1} = \theta, \quad (6)$$

where $\theta = \sigma_2 c_1^{1/2}/(\sigma_1 c_2^{1/2})$. However, the exact distribution of V given in Eq. 2 involves a beta mixture of noncentral t distributions. Thus, the associated properties can be notably different from a normal distribution for finite sample sizes. It is understandable that the particular identity of Eq. 6 will give a suboptimal result when the sample sizes are small. Such a phenomenon is demonstrated in the following illustration.

Total cost is fixed and actual power needs to be maximized

To develop a systematic search for the optimal solution, the aforementioned normal approximation is utilized as the benchmark in the exploration. It can be shown,under a fixed value of total cost C, that the maximum power is obtained with the sample size combination

$$N_{1Z} = \frac{C(\sigma_1 c_2^{1/2})}{c_1(\sigma_1 c_2^{1/2}) + c_2(\sigma_2 c_1^{1/2})} \text{ and } N_{2Z} = \frac{C(\sigma_2 c_1^{1/2})}{c_1(\sigma_1 c_2^{1/2}) + c_2(\sigma_2 c_1^{1/2})}. \quad (7)$$

It is easy to see that $c_1N_{1Z} + c_2N_{2Z} = C$ and $N_{2Z}/N_{1Z} = \theta$, as in Eq. 6. But in practice, the sample sizes need to be integer values, so the use of discrete numbers introduces some in exactness into the cost analysis. To find the proper result, a detailed power calculation and comparison are performed for the sample size combinations with $N_1$from $N_{1\min}$ to $N_{1\max}$ and $N_2 = \text{Floor}[(C - c_1N_1)/c_2]$, where $N_{1\min} = \text{Floor}(N_{1Z}) - 1$, $N_{1\max} = \text{Floor}[\{C - c_2$

Table 2 Computed sample sizes (N_1, N_2) and actual power when sample size N_2 is fixed with μ_d = 1, α = .05 and 1 –β = .90

σ_1:σ_2														
1/3:1			1/2:1			1:1			2:1			3:1		
N_1	N_2	Power	N_1	N_2	Power	N_1	N_2	Power	N_1	N_2	Power	N_1	N_2	Power
7	15	.9086	11	16	.9057	18	30	.9032	55	50	.9005	108	100	.9014
5	18	.9228	9	18	.9131	16	40	.9027	49	100	.9015	102	200	.9009
4	21	.9157	8	20	.9185	15	50	.9011	48	150	.9056	100	300	.9004

$(\text{Floor}(N_{2Z}) - 1)\}/c_1]$, and the function Floor(a) returns the largest integer that is less than or equal to a. Thus theoptimal sample size allocation is the one giving the largest power. Numerical results are given in Table 3 for (c_1, c_2) = (1, 1), (1, 2), and (1, 3) and fixed total cost C = 25, 30, 50, 100, and 180 in accordance with the five standard deviation settings of σ_1 and σ_2 reported in the previous two tables. Examination of the results in Table 3 reveals that the actual power for a given total cost deceases drasticallyas the unit cost c_2 increases from 1 to 3. Regarding the optimal allocation, the general formula for the sample size ratio presented in Eq. 6 does not hold in several cases. For example, the ratio $N_2/N_1 = 11/17 = 0.6471$ for (σ_1, σ_2) = (1, 1) and (c_1, c_2) = (1, 3) is slightly greater than the ratiocomputed with Eq. 6: $\theta = \left(1 \cdot 1^{1/2}\right)/\left(1 \cdot 3^{1/2}\right) = 0.5774$. It should be noted that Guo and Luh (2009, Eq. 20) give the same approximate sample size formulas as in Eq. 7. However, they did not discuss how to utilize the particular result to find the ideal sample sizes for a fixed cost. Also, the numerical demonstration of Guo and Luh (p. 291) did not provide a systematic search for the optimal solution, and the sample sizes reported in their exposition are not integers. Ultimately, the inexactness issue incurred by integer sample sizes in cost analysis is not addressed by Guo and Luh.

Target power is fixed and total cost needs to be minimized

In contrast to the previous situation where costs were fixed, the strategy to accommodate both power performance and cost appraisal can be conducted by finding the optimal allocation for minimizing cost when the target power is prechosen. In this case, the large-sample theory shows that in order to ensure the nominal power while minimizing total cost $C = c_1N_{1Z} + c_2N_{2Z}$, the best sample size combination is

$$N_{1Z} = \frac{\theta\sigma_1^2 + \sigma_2^2}{\theta\mu_d^2/\left(Z_{\alpha/2} + Z_\beta\right)^2} \text{ and } N_{2Z} = \frac{\theta\sigma_1^2 + \sigma_2^2}{\mu_d^2/\left(Z_{\alpha/2} + Z_\beta\right)^2}, \tag{8}$$

where θ is the optimal ratio in Eq. 6. It can be readily seen that $N_{2Z}/N_{1Z} = \theta$ and $\sigma_1^2/N_{1Z} + \sigma_2^2/N_{2Z} = \mu_d^2/\left(Z_{\alpha/2} + Z_\beta\right)^2$. Due to the discrete character of sample size, the optimal allocation is found through a screening of sample size combinations that attain the desired power while giving the least cost. The exact power computation and cost evaluation are conducted for sample size combinations with N_1 from $N_{1\min}$ to $N_{1\max}$ and a proper value of $N_2 \geq \text{Floor}\left[\sigma_2^2/\{\mu_d^2/(z_{\alpha/2} + z_\beta)^2 - \sigma_1^2/N_1\}\right]$ satisfying the required power, where $N_{1\min} = \text{Floor}(N_{1Z}), N_{1\max} = \text{Ceil}\left[\sigma_1^2/\{\mu_d^2/(z_{\alpha/2} + z_\beta)^2 - \sigma_2^2/(Floor(N_{2Z}) - 1)\}\right]$, and the function Ceil($a$) returns the smallest integer that is greater than or equal to a. Thus, the optimal sample size allocation is the one giving the smallest cost while maintaining the specified power level. In cases where there is more than one combination yielding the same magnitude of least cost, the one producing the larger power is reported. Table 4 provides the corresponding optimal sample size allocation, cost, and actual power for the configurations of (c_1, c_2) = (1, 1), (1, 2), and (1, 3) and the five standard deviation settings of σ_1 and σ_2 in the preceding tables. It is clear that the total cost for a required power and fixed standard deviations increases substantially as the unit cost c_2 changes from 1 to 3. Again, the sample size ratios are close to, but different from, the approximate ratio θ. The largest discrepancy occurs with the case $N_2/N_1 = 16/6 = 2.6667$ for (σ_1, σ_2) = (1/3, 1) and (c_1, c_2) = (1, 1), where as the counterpart ratio $\theta = (1 \cdot 1^{1/2})/(1/3 \cdot 1^{1/2}) = 3$.

To demonstrate the advantage and importance of the exact technique, we alsoexamine the theoretical and empirical properties of the approximate methods of Schouten (1999) and Singer (2001). Accordingly, Schouten's (p. 90) formulas are based on the normal approximation and give the identical approximate estimates N_{1Z} and N_{2Z} as defined in Eq. 8. In view of the approximate t distribution of the Welch's test statistic V defined in Eq. 1, Singer (Eq. 2) suggested a modification of Eq. 4 by replacing the percentiles of standard normal distribution with those of a t distribution with degrees

Table 3 Computed sample sizes (N_1, N_2) and actual power when the total cost is fixed with μ_d = 1 andα = .05

σ_1:σ_2	1/3:1				1/2:1				1:1				2:1				3:1			
c_1:c_2	Cost	N_1	N_2	Power	Cost	N_1	N_2	Power	Cost	N_1	N_2	Power	Cost	N_1	N_2	Power	Cost	N_1	N_2	Power
1:1	25	6	19	.9467	30	10	20	.9403	50	25	25	.9334	100	67	33	.9099	180	135	45	.9156
1:2	25	5	10	.7432	30	8	11	.7608	50	20	15	.8076	100	58	21	.8229	180	122	29	.8548
1:3	25	4	7	.5570	30	6	8	.5984	50	17	11	.6917	100	52	16	.7473	180	114	22	.8016

Table 4 Computed sample sizes (N_1, N_2), cost, and actual power when the total cost needs to be minimized with target power $1-\beta = .90$, $\mu_d = 1$, and $\alpha = .05$

σ_1:σ_2	1/3:1				1/2:1				1:1				2:1				3:1			
c_1:c_2	Cost	N_1	N_2	Power	Cost	N_1	N_2	Power	Cost	N_1	N_2	Power	Cost	N_1	N_2	Power	Cost	N_1	N_2	Power
1:1	22	6	16	.9144	26	9	17	.9017	45	23	22	.9057	97	65	32	.9013	171	128	43	.9015
1:2	37	7	15	.9086	43	11	16	.9057	65	27	19	.9020	126	74	26	.9015	208	140	34	.9009
1:3	51	9	14	.9014	58	13	15	.9012	84	30	18	.9032	151	79	24	.9015	239	149	30	.9003

of freedom $\hat{v}$. Specifically, it requires an iterative process to find the smallest integer that satisfies the inequality

$$N_{1S} \geq (\sigma_1^2 + \sigma_2^2/r_s)(t_{\hat{v},\alpha/2} + t_{\hat{v},\beta})^2/\mu_d^2, \tag{9}$$

where $r_S = N_{2S}/N_{1S}$. However, Singer did not provide any analytical justification for this alternative expression. Essentially, the naive formulation of Eq. 9 is questionable for lack of theoretical explanation. It is well known that if $Z \sim N(0, 1)$, then $X = (Z + \mu) \sim N(\mu, 1)$, where μ is a constant. This particular result and related properties yield the approximate formulas in Eq. 8. On the other hand, the linear transformation of the normal distribution does not generalize to the case of the t distribution; that is, if $t \sim t(df)$, then $Y = (t + \mu)$ does not follow a noncentral t distribution $t(df, \mu)$ with a noncentrality parameter μ and degrees of freedom df. Actually, a random variable Y is said to have a noncentral t distribution $t(df, \mu)$ if and only if $Y = (Z + \mu)/(W/df)^{1/2}$, where $Z \sim N(0, 1)$, $W \sim \chi^2(df)$, and Z and W are independent (Rencher, 2000, pp. 102–103). This may explain the fact that direct substitution of standard normal percentiles with those of t distribution was rarely described in the literature of sample size methodology. Instead, an iterative search is required to resolve the issue for statistical reasoning and exactness. Nevertheless, Guo and Luh (2009) applied Eq. 9 with $r_S = \theta$ to determine optimal sample sizes when target power is fixed and total cost needs to be minimized.

For the purpose of comparison, we performed an extensive numerical examination of sample size calculations for the model settings in Table 4 of Guo and Luh (2009). To our knowledge, no research to date has compared the performance of the available approximate procedures with the exact method. All the sample sizes, cost, and corresponding actual power of the two approximate methods of Schouten (1999) and Singer (2001) and the exact approach are presented in Table 5. For target power $1 - \beta = .80$, $\mu_d = 1$, and $\alpha = .05$, a total of 24 model settings are examined according to the combined configurations of standard deviation ratio (σ_1:σ_2 = 1: 1 and 1: 2) and unit cost ratio (c_1:c_2 = 1: 2, 1: 1, and 2: 3) for $\sigma_1^2 = 1.00$, 2.15, 1.46, and 4.18. The sample sizes computed by Schouten's method are denoted by N_{1Z} and N_{2Z}, whereas the sample sizes N_{1S} and N_{2S} listed in Table 5 for the procedure of Singer are exact replicates of those presented for the untrimmed case in Table 4 of Guo and Luh. The corresponding exact sample sizes computed with the suggested approach are expressed as N_{1E} and N_{2E}.

It can be readily seen from Table 5 that there are discrepancies between the approximate and exact procedures. First, the normal approximation or Schouten's (1999) method is misleading because only 4 out of 24 cases have attained the target power level of .80 (cases 4, 6, 12, and 24). Thus, the sample sizes N_{1Z} and N_{2Z} are generally inadequate. For the four occasions that meet the minimum power requirement, the resulting costs of cases 6, 12, and 24 are larger than those of the exact approach. Again, the reported sample sizes N_{1Z} and N_{2Z} are not optimal. Accordingly, case 4 is the single instance that agrees with the exact result. On the other hand, all the sample sizes N_{1S} and N_{2S} associated with Singer's (2001) method satisfy the necessary minimum power .80. While there are seven occurrences (cases 2, 4, 8, 14, 15, 19, and 20) that match the exact results, the other 17 sample size N_{1S} and N_{2S} combinations suffer the disadvantage of incurring higher cost than the optimal selections N_{1E} and N_{2E}. In view of these empirical evidences, it is clear that the existing approximate procedures of Schouten and Singer are not accurate enough to guarantee optimal sample sizes and, therefore, the procedures presented in Eqs. 8 and 9 are not recommended.

Furthermore, Lee (1992) examined the same problem-without considering the differential unit cost per subject in the two groups, and this can be viewed as a special case of the presentation here with $c_1 = c_2 = 1$. Accordingly, his algorithm for determining the optimal sample sizes is questionable. For example, when $\sigma_1 = \sigma_2 = 1$, the reported sample sizes are $N_1 = N_2 = 23$ with total cost = total sample size = 46, and actual power is .9121. In contrast, our computation gives $N_1 = 23$ and $N_2 = 22$, with total cost = total sample size = 45, and attained power is. 9057. Therefore, to maintain the least target power level of .90, it requires only a total of 45 sample sizes, rather than the sizes of 46 as reported by Lee. Consequently, it is worthwhile conducting the suggested exact sample size computations.

Table 5 Computed sample sizes (N_1, N_2), cost, and actual power for different procedures when the total cost needs to be minimized with target power $1-\beta = .80$, $\mu_d = 1$, and $\alpha = .05$

				Schouten				Singer				Exact Method			
Case	σ_1:σ_2	c_1:c_2	θ	N_{1Z}	N_{2Z}	Cost	Power	N_{1S}	N_{2S}	Cost	Power	N_{1E}	N_{2E}	Cost	Power
$\sigma_1^2 = 1$															
1	1:1	1:2	0.71	19	14	47	.7811	21	15	51	.8156	20	15	50	.8076
2	1:1	1:1	1.00	16	16	32	.7798	17	17	34	.8058	17	17	34	.8058
3	1:1	2:3	0.82	18	15	81	.7889	19	16	86	.8141	18	16	84	.8040
4	1:2	1:2	1.41	31	44	119	.8017	31	44	119	.8017	31	44	119	.8017
5	1:2	1:1	2.00	24	48	72	.7963	25	49	74	.8073	24	49	73	.8018
6	1:2	2:3	1.63	28	46	194	.8037	28	46	194	.8037	29	45	193	.8013
$\sigma_1^2 = 2.15$															
7	1:1	1:2	0.71	41	29	99	.7895	43	30	103	.8056	42	30	102	.8018
8	1:1	1:1	1.00	34	34	68	.7910	35	35	70	.8028	35	35	70	.8028
9	1:1	2:3	0.82	38	32	172	.7995	39	32	174	.8042	40	31	173	.8014
10	1:2	1:2	1.41	65	92	249	.7972	66	93	252	.8020	65	93	251	.8004
11	1:2	1:1	2.00	51	102	153	.7978	52	103	155	.8030	51	103	154	.8004
12	1:2	2:3	1.63	59	97	409	.8020	60	97	411	.8040	58	97	407	.8001
$\sigma_1^2 = 1.46$															
13	1:1	1:2	0.71	28	20	68	.7875	30	21	72	.8110	29	21	71	.8055
14	1:1	1:1	1.00	23	23	46	.7830	24	24	48	.8008	24	24	48	.8008
15	1:1	2:3	0.82	26	22	118	.7975	27	22	120	.8044	27	22	120	.8044
16	1:2	1:2	1.41	44	63	170	.7968	45	64	173	.8038	44	64	172	.8014
17	1:2	1:1	2.00	35	70	105	.7995	36	71	107	.8070	35	71	106	.8033
18	1:2	2:3	1.63	40	66	278	.7999	41	66	280	.8027	39	67	279	.8012
$\sigma_1^2 = 4.18$															
19	1:1	1:2	0.71	80	57	194	.7993	81	57	195	.8013	81	57	195	.8013
20	1:1	1:1	1.00	66	66	132	.7964	67	67	134	.8024	67	67	134	.8024
21	1:1	2:3	0.82	73	60	326	.7954	75	61	333	.8038	75	60	330	.8002
22	1:2	1:2	1.41	126	179	484	.7998	127	180	487	.8022	127	179	485	.8006
23	1:2	1:1	2.00	99	198	297	.7997	100	199	299	.8024	99	199	298	.8010
24	1:2	2:3	1.63	114	187	789	.8015	115	187	791	.8025	113	187	787	.8005

Numerical example

To demonstrate the features crossing different allocation constraints and cost considerations in sample size planning, the comparison of ability tests administered online and in the laboratory of Ihme et al. (2009) is used as an example. The test scores collected online and offline are assumed to have normal distributions with different variances, because the demographical structure of online samples can differ from that of offline samples acquired in conventional laboratory settings. To illustrate sample size determination for design planning, the results of Ihme et al. are modified to have the underlying population parameter values of $\mu_{\text{Lab}} = 11$, $\mu_{\text{Online}} = 10$, $\sigma_{\text{Lab}} = 2.3$, and $\sigma_{\text{Online}} = 2.7$. It is clear that online testing has the advantages of ease of obtaining a large sample and low cost. Thus, it may be desirable to set the sample ratio as $N_{\text{Online}}/N_{\text{Lab}} = 4/1$, which would imply that the sample sizes required to attain power .90 at the significance level .05 are $N_{\text{Lab}} = 76$ and $N_{\text{Online}} = 304$. In case in which the sample size N_{Online} is selected as 400, the offline group needs sample size $N_{\text{Lab}} = 71$ to meet the same power and significance requirements. However, it is important to take budget issues into account. Assume that the available total cost is set as $C = 100$ and the respective unit costs per subject are $c_{\text{Lab}} = 1$ and $c_{\text{Online}} = 0.2$. The optimal sample size solutionis $N_{\text{Lab}} = 65$ and $N_{\text{Online}} = 175$, which has an actual power of .8079. On the other hand, to attain the pre-assigned power of .90, the design must have the sample size allocation as $N_{\text{Lab}} = 86$ and $N_{\text{Online}} = 224$, which amounts to the budget of C =130.8. Such information may be useful for investigators to justify the design strategy and financial support. Although they did not address the sample size calculation, the reader is referred to Ihme et al. for further details about online achievement tests.

Conclusion

The problem of testing the equality of the means from two independent and normally distributed populations with unknown and unequal variances has beenwidely considered in the literature. The distinctive usefulness of Welch's (1938) test in applications further occasions methodological and practical concerns about the corresponding procedures for sample size determination. Computationally, the use of computers and the general availability of statistical software permit inherent requirements for exact analysis. In view of the importance of sample size calculations in actual practice and the limited features of available computer packages, the corresponding programs are developed to facilitate the usage of the suggested approaches. Intensive numerical integration and incremental search are incorporated in the presented computer algorithms for finding the optimal solutions for different design requirements. Furthermore, various sample size tables are provided to help researchers have a better understanding of the inherent relationship that exists between the planned sample sizes conditional on the model configurations. The proposed sample size procedures enhance and expand the current methods and should be useful for planning of research in two-group situations where variances and costs per subject both differ across groups.

Author Note The authors thank the editor, Gregory Francis, and the three anonymous reviewers for their helpful comments. This research was partially supported by National Science Council Grant NSC-99-2118-M-033-002.

References

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.

Bacchetti, P. (2010). Current sample size conventions: Flaws, harms, and alternatives. *BMC Medical Research Methodology, 8*, 17.

Bacchetti, P., McCulloch, C. E., & Segal, M. R. (2008). Simple, defensible sample sizes based on cost efficiency. *Biometrics, 64*, 577–594.

Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine, 14*, 1933–1940.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

R Development Core Team (2010). *R: A language and environment for statistical computing* [Computer software and manual]. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Fleiss, J. L., Tytun, A., & Ury, H. K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics, 36*, 343–346.

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

Guo, J. H., & Luh, W. M. (2009). Optimum sample size allocation to minimize cost or maximize power for the two-sample trimmed mean test. *The British Journal of Mathematical and Statistical Psychology, 62*, 283–298.

Heilbrun, L. K., & McGee, D. L. (1985). Sample size determination for the comparison of normal means when one sample size is fixed. *Computational Statistics and Data Analysis, 3*, 99–102.

Ihme, J. M., Lemke, F., Lieder, K., Martin, F., Muller, J. C., & Schmidt, S. (2009). Comparison of ability tests administered online and in the laboratory. *Behavior Research Methods, 41*, 1183–1189.

Kieser, M., & Wassmer, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal, 38*, 941–949.

Kim, S. H., & Cohen, A. S. (1998). On the Behrens–Fisher problem: A review. *Journal of Educational and Behavioral Statistics, 23*, 356–377.

Lee, A. F. S. (1992). Optimal sample sizes determined by two-sample Welch's *t* test. *Communications in Statistics—Simulation and Computation, 21*, 689–696.

Lee, A. F. S., & Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. *Journal of the American Statistical Association, 70*, 933–941.

Luh, W. M., & Guo, J. H. (2007). Approximate sample size formulas for the two-sample trimmed mean test with unequal variances. *The British Journal of Mathematical and Statistical Psychology, 60*, 137–146.

Luh, W. M., & Guo, J. H. (2010). The sample size needed for the trimmed *t* test when one group size is fixed. *Journal of Experimental Education, 78*, 14–25.

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Mahwah, NJ: Erlbaum.

Nam, J. M. (1973). Optimum sample sizes for the comparison of the control and treatment. *Biometrics, 29*, 101–108.

Nel, D. G., van der Merwe, C. A., & Moser, B. K. (1990). The exact distribution of the univariate and multivariate Behrens–Fisher statistics with a comparison of several solutions in the univariate case. *Communications in Statistics—Theory and Methods, 19*, 279–298.

Rencher, A. C. (2000). *Linear models in statistics*. New York: Wiley.

SAS Institute. (2008a). *SAS/IML user's guide, version9.2*. Cary, NC: SAS Institute Inc.

SAS Institute. (2008b). *SAS/STATuser's guide, version9.2*. Cary, NC: SAS Institute Inc.

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

Schouten, H. J. A. (1999). Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine, 18*, 87–91.

Singer, J. (2001). A simple procedure to compute the sample size needed to compare two independent groups when the population variances are unequal. *Statistics in Medicine, 20*, 1089–1095.

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.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismalia, A., Rios, L. P., et al. (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology, 10*, 1.

Wang, Y. Y. (1971). Probabilities of the type I errors of the Welch tests for the Behrens–Fisher problem. *Journal of the American Statistical Association, 66*, 605–608.

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