PSYCHOMETRIKA
https://doi.org/10.1007/s11336-019-09692-3

POWER ANALYSIS AND SAMPLE SIZE PLANNING IN ANCOVA DESIGNS

GWOWEN SHIEH
NATIONAL CHIAO TUNG UNIVERSITY

The analysis of covariance (ANCOVA) has notably proven to be an effective tool in a broad range of scientific applications. Despite the well-documented literature about its principal uses and statistical properties, the corresponding power analysis for the general linear hypothesis tests of treatment differences remains a less discussed issue. The frequently recommended procedure is a direct application of the ANOVA formula in combination with a reduced degrees of freedom and a correlation-adjusted variance. This article aims to explicate the conceptual problems and practical limitations of the common method. An exact approach is proposed for power and sample size calculations in ANCOVA with random assignment and multinormal covariates. Both theoretical examination and numerical simulation are presented to justify the advantages of the suggested technique over the current formula. The improved solution is illustrated with an example regarding the comparative effectiveness of interventions. In order to facilitate the application of the described power and sample size calculations, accompanying computer programs are also presented.

Key words: general linear hypothesis, omnibus test, power, sample size.

1. Introduction

The analysis of covariance (ANCOVA) was originally developed by Fisher (1932) to reduce error variance in experimental studies. Its essential nature and principal use were well explicated by Cochran (1957) and subsequent articles in the same issue of *Biometrics*. The value and use of ANCOVA have also received considerable attention in social science, for example, see Elashoff (1969), Keselman et al. (1998), and Porter and Raudenbush (1987). Comprehensive introduction and fundamental principles can be found in the excellent texts of Fleiss (2011), Huitema (2011), Keppel and Wickens (2004), Maxwell and Delaney (2004), and Rutherford (2011). It is essential to note that ANCOVA provides a useful approach for combining the advantages of two highly acclaimed procedures of analysis of variance (ANOVA) and multiple linear regression. The extensive literature shows that it is one of the major methods of statistical analysis in applied research across many scientific fields.

The importance and implications of statistical power analysis in scientific research are well demonstrated in Cohen (1988), Kraemer and Blasey (2015), Murphy et al. (2014), and Ryan (2013), among others. Accordingly, it is of great practical value to develop theoretically sound and numerically accurate power and sample size procedures for detecting treatment differences within the context of ANCOVA. There are numerous published sources that address statistical theory and applications of power analysis for ANOVA and multiple linear regression. Specifically, various algorithms and tables for power and sample size calculations in ANOVA have been presented in the classic sources of Bratcher et al. (1970), Pearson and Hartley (1951), Scheffe (1961), and Tiku (1967, 1972). The corresponding results for multiple regression and correlation, especially the distinct notion of fixed and random regression settings, were given in Gatsonis

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-019-09692-3.

Correspondence should be made to Gwowen Shieh, Department of Management Science, National Chiao Tung University, Hsinchu 30010, Taiwan, ROC. Email: gwshieh@mail.nctu.edu.tw

© 2019 The Author(s), corrected publication 2021
Published online: 10 December 2019

and Sampson (1989), Mendoza and Stafford (2001), Sampson (1974), and Shieh (2006, 2007). However, relatively little research has attempted to address the corresponding issues for ANCOVA.

This lack of further discussion can partly be attributed to the simple framework and conceptual modification of Cohen (1988) on the use of ANOVA method for power evaluation in ANCOVA research. It is argued that the ANCOVA of original responses is essentially the ANOVA of the regression-adjusted or statistically controlled measurements obtained from the linear regression of unadjusted responses on the covariates that is common to all treatment groups. However, some modifications are required to account for the number of covariate variables and the strength of correlation between the response and covariate variables. Accordingly, both the error degrees of freedom and variance component are reduced. Then, the power and sample size computations in ANCOVA proceed in exactly the same way as in analogous ANOVA designs. The methodology of Cohen (1988) has become common practice for power analysis in ANCOVA settings as repeatedly demonstrated in Huitema (2011), Keppel and Wickens (2004), Levin (1997), Maxwell and Delaney (2004), and Yang et al. (1996).

It is well known that the ANOVA adopts the fundamental assumptions of independence, normality, and constant variance. The corresponding hypothesis testing and theoretical considerations are valid only if these assumptions are satisfied. The consequences of violations of independence assumption in ANOVA have been reported in Kenny and Judd (1986), Pavur and Nath (1984), and Scariano and Davenport (1987), among others. An essential assumption underlying ANCOVA is the regression coefficients associating the response variable with the covariate variables are the same for each treatment group. Therefore, the regression adjustment in Cohen's (1988, pp. 379–380) covariance framework includes the common regression coefficient estimates derived from the multiple regression between the response and covariate variables across all treatment groups. Unlike the original responses, the adjusted responses are generally correlated and thus violate the independence of observations assumption for ANOVA. Therefore, Cohen's (1988) procedure is intrinsically inexact, even with the technical considerations of a deflated degrees of freedom and a correlation-adjusted variance. Consequently, this prevailing method only provides approximate power and sample size calculations in ANCOVA designs. It should be stressed that no research to date has acknowledged this crucial problem and the result has most likely been interpreted as an exact solution.

Toward the goal of choosing the most appropriate methodology for ANCOVA studies, the present article focuses on the Wald tests for the general linear hypothesis of treatment effects. Under the two different assumptions of a priori specified covariate values and multinormal distributed covariate variables, the exact power functions of the Wald statistic are derived. The analytic derivations for a general linear hypothesis require the involved operations of matrix algebra and sophisticated evaluations of matrix t variables that have not been reported elsewhere. Detailed numerical investigations were conducted to evaluate the existing formulas for power and sample size computations under a wide range of model settings, including non-normal covariate variables. According to the analytic justification and empirical assessment, the suggested approach has a decisive advantage over the conventional method. An applied example regarding the comparative effectiveness of interventions is presented to illustrate the distinct features and practical usefulness of the proposed techniques. Computer codes are also presented to implement the recommended power calculation and sample size determination in planning ANCOVA studies.

2. General Linear Hypothesis

A one-way fixed-effects ANCOVA model with multiple covariates can be expressed as

$$Y_{ij} = \mu_i + \sum_{k=1}^{P} X_{kij}\beta_k + \varepsilon_{ij}, \tag{1}$$

where Y_{ij} is the score of the jth subject in the ith treatment group on the response variable, μ_i is the ith group intercept, X_{kij} is the score of the jth subject in the ith treatment group on the kth covariate, β_k is the slope coefficient of the kth covariate, and ε_{ij} is the independent $N(0, \sigma^2)$ error with $i = 1, \ldots, G\ (\geq 2)$, $j = 1, \ldots, N_i$, and $k = 1, \ldots, P\ (\geq 1)$. The least-square estimator for the ith intercept μ_i is given by

$$\hat{\mu}_i = \bar{Y}_i - \sum_{k=1}^{P} \bar{X}_{ki\cdot}\hat{\beta}_k, \tag{2}$$

where $\bar{Y}_i = \sum_{j=1}^{N_i} Y_{ij}/N_i$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_P)^{\mathrm{T}} = \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}$, $\mathbf{S}_{XX} = \sum_{i=1}^{G}\sum_{j=1}^{Ni}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^{\mathrm{T}}$, $\mathbf{S}_{XY} = \sum_{i=1}^{G}\sum_{j=1}^{Ni}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(Y_{ij} - \bar{Y}_i)$, $\mathbf{X}_{ij} = (X_{1ij}, \ldots, X_{Pij})^{\mathrm{T}}$, $\bar{\mathbf{X}}_i = \sum_{j=1}^{Ni}\mathbf{X}_{ij}/N_i = (\bar{X}_{1i\cdot}, \ldots, \bar{X}_{Pi\cdot})^{\mathrm{T}}$, and $\bar{X}_{ki\cdot} = \sum_{j=1}^{Ni} X_{kij}/N_i$. Accordingly, the least-squares estimators $\hat{\mu}_i$ of μ_i have the following distributions:

$$\hat{\mu}_i \sim N(\mu_i, \sigma^2\{1/N_i + \bar{\mathbf{X}}_i^{\mathrm{T}}\mathbf{S}_{XX}^{-1}\bar{\mathbf{X}}_i\}) \text{ and } Cov(\hat{\mu}_i, \hat{\mu}_{i'}) = \sigma^2\bar{\mathbf{X}}_i^{\mathrm{T}}\mathbf{S}_{XX}^{-1}\bar{\mathbf{X}}_{i'} \tag{3}$$

for $i \neq i'$, i and $i' = 1, \ldots, G$. Because the covariances between regression-adjusted estimators $\{\hat{\mu}_1, \ldots, \hat{\mu}_G\}$ are generally not zero, they should not be treated as independent variables. For notational simplicity, the prescribed properties are expressed in matrix form:

$$\hat{\boldsymbol{\mu}} \sim N_G(\boldsymbol{\mu}, \sigma^2\mathbf{V}), \tag{4}$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_G)^{\mathrm{T}}, \boldsymbol{\mu} = (\mu_1, \ldots, \mu_G)^{\mathrm{T}}, \mathbf{V} = \mathbf{D} + \bar{\mathbf{X}}^{\mathrm{T}}\mathbf{S}_{XX}^{-1}\bar{\mathbf{X}}, \mathbf{D} = \mathrm{Diag}(1/N_1, \ldots, 1/N_G)$ is the $G \times G$ diagonal matrix with diagonal elements $\{1/N_1, \ldots, 1/N_G\}$, and $\bar{\mathbf{X}} = (\bar{\mathbf{X}}_1, \ldots, \bar{\mathbf{X}}_G)$.

The adjusted group means are the expected group responses evaluated at the grand covariate means:

$$\mu_i^* = \mu_i + \sum_{k=1}^{P} \bar{X}_{k\cdot\cdot}\beta_k \quad \text{for } i = 1, \ldots, G, \tag{5}$$

where $\bar{X}_{k\cdot\cdot} = \sum_{i=1}^{G}\sum_{j=1}^{Ni} X_{kij}/N_T, k = 1, \ldots, P$, and $N_T = \sum_{i=1}^{G} N_i$. A natural and unbiased estimator of the adjusted group mean μ_i^* is

$$\hat{\mu}_i^* = \hat{\mu}_i + \sum_{k=1}^{P} \bar{X}_{k\cdot\cdot}\hat{\beta}_k = \bar{Y}_i - \sum_{k=1}^{P} \hat{\beta}_k(\bar{X}_{ki\cdot} - \bar{X}_{k\cdot\cdot}). \tag{6}$$

Then, the least-squares estimators $\hat{\mu}_i^*$ of the adjusted group means μ_i^* have the following distributions:

$$\hat{\mu}_i^* \sim N(\mu_i^*, \sigma^2\{1/N_i + (\bar{\mathbf{X}}_i - \mathbf{M})^{\mathrm{T}}\mathbf{S}_{XX}^{-1}(\bar{\mathbf{X}}_i - \mathbf{M})\})$$

and

$$Cov(\hat{\mu}_i, \hat{\mu}_{i'}) = \sigma^2(\bar{\mathbf{X}}_i - \mathbf{M})^{\mathrm{T}}\mathbf{S}_{XX}^{-1}(\bar{\mathbf{X}}_{i'} - \mathbf{M}), \tag{7}$$

where $\mathbf{M} = \sum_{i=1}^{G}\sum_{j=1}^{Ni}\mathbf{X}_{ij}/N_T = (\bar{X}_{1\cdot\cdot}, \ldots, \bar{X}_{P\cdot\cdot})^{\mathrm{T}}$ for $i \neq i'$, i and $i' = 1, \ldots, G$. The vector of adjusted group mean estimators $\hat{\boldsymbol{\mu}}^* = (\hat{\mu}_1^*, \ldots, \hat{\mu}_G^*)^{\mathrm{T}}$ has the distribution

$$\hat{\boldsymbol{\mu}}^* \sim N_G(\boldsymbol{\mu}^*, \sigma^2\mathbf{V}^*), \tag{8}$$

where $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_G^*)^{\mathrm{T}}$, $\mathbf{V}^* = \mathbf{D} + (\bar{\mathbf{X}} - \mathbf{M}\mathbf{1}_G^{\mathrm{T}})^{\mathrm{T}}\mathbf{S}_{XX}^{-1}(\bar{\mathbf{X}} - \mathbf{M}\mathbf{1}_G^{\mathrm{T}})$, and $\mathbf{1}_G$ is a $G \times 1$ column vector of all 1's.

To test the general linear hypothesis about treatment effects or adjusted mean effects in terms of

$$\mathrm{H}_0\colon \mathbf{C}\mu^* = \mathbf{0}_c \text{ versus } \mathrm{H}_1\colon \mathbf{C}\mu^* \neq \mathbf{0}_c, \tag{9}$$

where $\mathbf{C}$ is a $c \times G$ contrast matrix of full row rank and $\mathbf{0}_c$ is a $c \times 1$ null column vector, the Wald test statistic is of the form

$$W^* = (\mathbf{C}\hat{\boldsymbol{\mu}}^*)^{\mathrm{T}}(\mathbf{C}\mathbf{V}^*\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\mu}}^*)/\{(G-1)\hat{\sigma}^2\} \tag{10}$$

where $\hat{\sigma}^2 = SSE/\nu$, $SSE = \sum_{i=1}^{G}\sum_{j=1}^{Ni}(Y_{ij} - \bar{Y}_i)^2 - \mathbf{S}_{XY}^{\mathrm{T}}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}$, and $\nu = N_T - G - P$. Note that the contrast matrix is confined to satisfy $\mathbf{C}\mathbf{1}_G = \mathbf{0}_c$. Hence, the general linear hypothesis of $\mathrm{H}_0\colon \mathbf{C}\mu^* = \mathbf{0}_c$ versus $\mathrm{H}_1\colon \mathbf{C}\mu^* \neq \mathbf{0}_c$ is equivalent to

$$\mathrm{H}_0\colon \mathbf{C}\boldsymbol{\mu} = \mathbf{0}_c \text{ versus } \mathrm{H}_1\colon \mathbf{C}\boldsymbol{\mu} \neq \mathbf{0}_c. \tag{11}$$

Also, the Wald test statistic can be rewritten as

$$W^* = (\mathbf{C}\hat{\boldsymbol{\mu}})^{\mathrm{T}}(\mathbf{C}\mathbf{V}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\mu}})/\{(G-1)\hat{\sigma}^2\}. \tag{12}$$

The Wald-type test has great practical and pedagogical appeal than the test procedure under the full-reduced-model formulation. Because of its simplicity and generality, the associated properties are derived and presented in the subsequent illustration. Under the null hypothesis with $\mathbf{C}\boldsymbol{\mu} = \mathbf{0}_c$, the test statistic W^* has an F distribution

$$W^* \sim F(c, \nu), \tag{13}$$

where $F(c, \nu)$ is an F distribution with c and ν degrees of freedom, $\nu = N_T - G - P$, and $N_T = \sum_{i=1}^{G} N_i$. Hence, H_0 is rejected at the significance level α if $W^* > F_{c,\nu,\alpha}$, where $F_{c,\nu,\alpha}$ is the upper $(100 \cdot \alpha)$th percentile of the F distribution $F(c, \nu)$. For fixed covariate values of $\{\mathbf{X}_{ij}, j = 1, \ldots, N_i \text{ and } i = 1, \ldots, G\}$, the test statistic W^* has the general distribution

$$W^* \sim F(c, \nu, \Lambda), \tag{14}$$

where $F(c, \nu, \Lambda)$ is a non-central F distribution with c and ν degrees of freedom and non-centrality parameter

$$\Lambda = (\mathbf{C}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{C}\mathbf{V}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\boldsymbol{\mu})/\sigma^2. \tag{15}$$

The associated power function of the general linear hypothesis is readily obtained as

$$\Psi(\Lambda) = P\{F(c, \nu, \Lambda) > F_{c,\nu,\alpha}\}. \tag{16}$$

3. Random Covariate Models

The prescribed statistical inferences about the general linear hypothesis are based on the conditional distribution of the covariate outcomes. As noted in Gatsonis and Sampson (1989), Mendoza and Stafford (2001), and Sampson (1974), the actual values of covariates cannot be known in advance just as the primary responses. It is vital to treat the covariates as random variables and to derive the distribution of the test statistic over possible values of the covariate variables. Moreover, Elashoff (1969) and Harwell (2003) emphasized that the statistical assumptions underlying the ANCOVA include the random assignment of subjects to treatments and the covariate variables are independent of the treatment effects. Moreover, the normal covariate setting is commonly employed to provide a fundamental framework for analytical derivation and theoretical discussion in ANCOVA studies as in Elashoff (1969) and Harwell (2003). Thus, it is constructive to assume the covariates have independent and identical normal distribution

$$\mathbf{X}_{ij} \sim N_P(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \tag{17}$$

where $\boldsymbol{\theta}$ is a $P \times 1$ vector and $\boldsymbol{\Sigma}$ is a $P \times P$ positive-definite variance–covariance matrix for $i = 1, \ldots, G$, and $j = 1, \ldots, N_i$.

Under the multinormal distribution of $\{\mathbf{X}_{ij} \sim N_P(\boldsymbol{\theta}, \boldsymbol{\Sigma}),\ j = 1, \ldots, N_i \text{ and } i = 1, \ldots, G\}$, it is straightforward to show (Gupta and Nagar 1999, Theorem 2.3.10 and Theorem 3.3.6) that $\mathbf{Z} = \bar{\mathbf{X}}\mathbf{C}^{\mathrm{T}}(\mathbf{C}\mathbf{D}\mathbf{C}^{\mathrm{T}})^{-1/2}$ has a matrix normal distribution and $\mathbf{S}_{XX}$ has a Wishart distribution

$$\mathbf{Z} \sim N_{P,c}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_c) \text{ and } \mathbf{S}_{XX} \sim W_P(N_T - G, \boldsymbol{\Sigma}), \tag{18}$$

where $\mathbf{I}_c$ is an identity matrix of dimension c. Accordingly, both $\mathbf{T} = \{\mathbf{S}_{XX} + \mathbf{Z}\mathbf{Z}^{\mathrm{T}}\}^{-1/2}\mathbf{Z}$ and $\mathbf{T}^* = \mathbf{T}\boldsymbol{\xi}$ have an inverted matrix variate t-distribution (Gupta and Nagar 1999, Section 4.4):

$$\mathbf{T} \sim IT_{P,C}(\nu + 1, \mathbf{0}, \mathbf{I}_P, \mathbf{I}_c) \text{ and } \mathbf{T}^* \sim IT_{P,1}(\nu + 1, \mathbf{0}_P, \mathbf{I}_P, \Gamma), \tag{19}$$

where $\boldsymbol{\xi} = (\mathbf{C}\mathbf{D}\mathbf{C}^{\mathrm{T}})^{-1/2}(\mathbf{C}\boldsymbol{\mu})/\sigma$ and $\Gamma = \boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi} = (\mathbf{C}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{C}\mathbf{D}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\boldsymbol{\mu})/\sigma^2$. Moreover, $A^* = \mathbf{T}^{*\mathrm{T}}\mathbf{T}^*/\Gamma$ has a matrix variate beta type I distribution (Gupta and Nagar 1999, Theorem 5.2.4) or a Beta distribution

$$A^* \sim B_1^I(P/2, (\nu + 1)/2) \equiv \text{Beta}\{P/2, (\nu + 1)/2\}. \tag{20}$$

Following these results, standard matrix algebra shows that non-centrality parameter Λ defined in Equation 15 has the alternative form

$$\Lambda = \boldsymbol{\xi}^{\mathrm{T}}(\mathbf{I}_c - \mathbf{T}^{\mathrm{T}}\mathbf{T})\boldsymbol{\xi} = \Gamma B^*, \tag{21}$$

where $B^* = (1 - A^*) \sim \text{Beta}\{(\nu + 1)/2, P/2\}$. In connection with the effect size measures in ANOVA, the first component $\mathbf{\Gamma}$ in Λ is rewritten as

$$\Gamma = N_T \gamma^2 \tag{22}$$

where $\gamma^2 = \sigma_\gamma^2/\sigma^2$, $\sigma_\gamma^2 = (\mathbf{C}\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{C}\mathbf{Q}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\boldsymbol{\mu})$, $\mathbf{Q} = \text{Diag}(1/q_1, \ldots, 1/q_G)$, $q_i = N_i/N_T$ for $i = 1, \ldots, G$. Consequently, the non-centrality term Λ has a useful formulation

$$\Lambda = N_T \gamma^2 B^*. \tag{23}$$

It should be pointed out that Gupta and Nagar (1999) only provides the generic definition and analytic properties of an inverted matrix variate t-distribution. Their results are applied and extended here to the context of ANCOVA. Accordingly, under the random covariate modeling framework, the W^* statistic has the two-stage distribution

$$W^*|B^* \sim F(c, \nu, \Lambda) \text{ and } B^* \sim Beta\{(\nu + 1)/2, P/2\}. \tag{24}$$

The exact power function can be formulated as

$$\Psi_E(\Lambda) = E_B[P\{F(c, \nu, \Lambda) > F_{c, \nu, \Lambda}\}], \tag{25}$$

where the expectation E_B is taken with respect to the distribution of B^*.

Notably, the omnibus test of the equality of treatment effects is a special case of the general linear hypothesis by specifying the contrast matrix as $\mathbf{C}_D\boldsymbol{\mu} = \mathbf{0}_{(G-1)}$ where

$$\mathbf{C}_D = (\mathbf{1}_{(G-1)}, -\mathbf{I}_{(G-1)}) \tag{26}$$

is a $(G - 1) \times G$ contrast matrix of full row rank. The component γ^2 in the non-centrality term Λ is simplified as

$$\delta^2 = \sigma_\delta^2/\sigma^2, \tag{27}$$

where $\sigma_\delta^2 = \sum_{i=1}^G q_i(\mu_i - \tilde{\mu})^2$ and $\tilde{\mu} = \sum_{i=1}^G q_i\mu_i$. The corresponding non-central component Λ is expressed as

$$\Lambda_D = N_T \delta^2 B^*. \tag{28}$$

The power function of the omnibus F test of treatment differences is simplified as

$$\Psi_E(\Lambda_D) = E_B[P\{F(G - 1, \nu, \Lambda_D) > F_{(G-1), \nu, \alpha}\}]. \tag{29}$$

Note that σ_δ^2 reduces to the form $\sigma_\delta^2 = \sum_{i=1}^{G}(\mu_i - \bar{\mu})^2/G$ with $\bar{\mu} = \sum_{i=1}^{G}\mu_i/G$ when $q_i = 1/G$ for all $i = 1, \ldots, G$. Hence, δ^2 has the same form as the signal to noise ratio f^2 in ANOVA (Fleishman 1980) for balanced designs. Although the prescribed application of general linear hypothesis is discussed only from the perspective of a one-way ANCOVA design, the number of groups G may also represent the total number of combined factor levels of a multi-factor ANCOVA design. Hence, using a contrast matrix associated with a specific designated hypothesis, the same concept and process of assessing treatment effects can be readily extended to two-way and higher-order ANCOVA designs.

4. Sample Size Determination

It is essential to note that the power function Ψ_E depends on the group intercepts $\{\mu_1, \ldots, \mu_G\}$ and variance component σ^2 through the non-centrality Λ or the effect size γ^2, but not the covariate coefficients $\{\beta_1, \ldots, \beta_P\}$. Also, under the prescribed stochastic assumptions for the covariate variables, the multivariate normal distribution leads to the unique conditional property on a beta distribution in the general distribution of the test statistic W^*. Due to the fundamental property of the contrast matrix, the resulting distribution and power function do not depend on the mean vector $\boldsymbol{\theta}$ and variance–covariance matrix $\boldsymbol{\Sigma}$ of the multinormal covariate distribution. To determine sample sizes in planning research designs, the power functions Ψ_E can be applied to calculate the sample sizes $\{N_{E1}, \ldots, N_{EG}\}$ needed to attain the specified power $1 - \beta$ for the chosen significance level α, contrast matrix $\mathbf{C}$, intercept parameters $\{\mu_1, \ldots, \mu_G\}$, variance component σ^2, and the number of covariates P.

For an ANCOVA design with a priori designated sample size ratios $\{r_1, \ldots, r_G\}$ with $r_i = N_i/N_1$ for $i = 1, \ldots, G$. The required computation is simplified to deciding the minimum sample sizes N_{E1} (with $N_{Ei} = N_{E1} \cdot r_i, i = 2, \ldots, G$) required to achieve the selected power level with the power functions Ψ_E. Using the embedded functions in popular software systems, optimal sample sizes can be readily computed through an iterative process. The SAS/IML (SAS Institute 2017) and R (R Development Core Team 2017) programs employed to perform the suggested power and sample size calculations are available as supplementary material. The proposed power and sample size procedures for the general linear hypothesis tests of ANCOVA subsume the results in Shieh (2017) for a single contrast test as a special case. Notably, the derivations and manipulations of an inverted matrix variate t are more involved than that of a Hotelling's T^2 distribution as demonstrated in Shieh (2017).

Alternatively, a simple procedure for the comparison of treatment effects has been described in Cohen (1988, pp. 379–380). Unlike the proposed two-stage distribution, it is suggested that W^* has a simplified F distribution

$$W^* \sim F(G - 1, \nu, \Lambda_A), \tag{30}$$

where $\Lambda_A = N_T \delta^2$. The corresponding power function is of the form

$$\Psi_A(\Lambda_A) = P\{F(G - 1, \nu, \Lambda_A) > F_{(G-1),\,\nu,\,\alpha}\}. \tag{31}$$

It is easily seen from the model assumption given in Equation 1 that $\sigma_Y^2 = \mathrm{Var}(Y_{ij}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2$ and $\rho = Corr(Y_{ij}, \sum_{k=1}^{P} X_{kij}\beta_k) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/\{\sigma_Y^2 \cdot \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}\}^{1/2}$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^{\mathrm{T}}$. Hence, the advantage of ANCOVA over ANOVA in the reduction of error variance from σ_Y^2 to $\sigma^2 = (1-\rho^2)\sigma_Y^2$ by a factor $(1 - \rho^2)$. For ease of illustration, the power function of the omnibus F test of treatment differences in ANOVA is also presented here:

$$\Psi_O(\Lambda_O) = P\{F(G - 1, N_T - G, \Lambda_O) > F_{(G-1),\,(N_T-G),\,\alpha}\}, \tag{32}$$

where $\Lambda_O = (1 - \rho^2)\Lambda_A$. With the reduction of error variance from σ_Y^2 to $\sigma^2 = (1 - \rho^2)\sigma_Y^2$, it is evident that $\Lambda_O \leq \Lambda_A$. Hence, the computed power Ψ_O is generally less than Ψ_A when all other factors are fixed despite the marginal difference between the two error degrees of freedom $N_T - G$ and $\nu = N_T - G - P$.

The prevailing procedure of Cohen (1988) provides a direct application of the ANOVA formula in combination with a reduced degrees of freedom and a correlation-adjusted variance. It is computationally simple because the simple formulation Ψ_A depends only on a non-central F distribution. On the other hand, one critical disadvantage of this method is that the F distribution and the associated sample size formula do not fully take into account the distributional features of covariates. A direct comparison of the two non-centrality components in Equations 28 and 30 reveals that $\Lambda_D < \Lambda_A$ because $0 < B^* < 1$. This indicates that the power function Ψ_A tends to over-estimate the true power Ψ_E and it also leads to an under-estimated sample size in attaining a desired power level. Notably, the suggested exact procedure is of pedagogical importance and involves a beta mixture of non-central F distributions. These theoretical examinations assure that the proposed technique has analytical superiority over the current method of Cohen (1988). Their practical accuracy will be demonstrated in the succeeding empirical assessments.

5. Numerical Assessments

To further demonstrate the contrasting features and practical consequences of the proposed approach and existing methods, detailed empirical appraisals are conducted to examine their performance in power and sample size calculations. For ease of comparison, the numerical illustration considered in Maxwell and Delaney (2004, pp. 441-443) for sample size planning and power analysis is utilized as the fundamental framework.

In particular, Maxwell and Delaney (2004) described an ANOVA design with $G = 3$, group intercepts $\{\mu_1, \mu_2, \mu_3\} = \{400, 450, 500\}$, and error variance $\sigma_Y^2 = 10,000$. Then, an ANCOVA model is introduced with the inclusion of an influential covariate variable X with $\rho = Corr(X, Y) = 0.5$ to partially account for the variance in the response variable Y. The corresponding unexplained error variance σ^2 in ANCOVA is reduced as $\sigma^2 = (1 - \rho^2)\sigma_Y^2 = 7,500$. To detect the treatment differences, they showed that the total sample sizes required to have a nominal power of 0.80 are 63 and 48 for the balanced ANOVA and ANCOVA designs, respectively. Thus, the ANCOVA design has the potential benefits to attain the same power with nearly 25% fewer subjects than an ANOVA. It should be noted that the power formulas Ψ_A and Ψ_O given in Equations 31 and 32, respectively, were applied for sample size calculations in Maxwell and Delaney (2004). To show a profound implication of the sample size procedures, extensive simulation study was performed under a wide range of model configurations.

First, the number of covariates and the population correlation between the response and covariate variables are extended to $P = 1, \ldots, 10$ and $\rho = 0.1$, 0.5, and 0.9. In each combined case of P and ρ, the required total sample sizes N_{TO}, N_{TA}, and N_{TE} are computed with the power functions Ψ_O, Ψ_A and Ψ_E for the ANOVA, approximate ANCOVA, and exact ANCOVA methods, respectively. Throughout this numerical investigation, the significance level and nominal power are chosen as $\alpha = 0.05$, and $1-\beta = 0.80$, respectively. Note that the effect sizes associated with $\rho = 0.1$, 0.5, and 0.9 are $\delta^2 = 0.1684$, 0.2222, and 0.8772, respectively. Second, to assess the potential impact of different and smaller effect sizes, the intercept parameters are modified as $\{\mu_1, \mu_2, \mu_3\} = \{410, 450, 490\}$ in the second set of numerical investigations. The resulting effect sizes are $\delta^2 = 0.1077$, 0.1422, and 0.5614 for $\rho = 0.1$, 0.5, and 0.9, respectively. Overall, these considerations result in a total of 60 different combined configurations. For $\{\mu_1, \mu_2, \mu_3\} = \{400, 450, 500\}$, the computed total sample sizes N_T are summarized in Tables 1, 2 and 3 for $\rho = 0.1, 0.5$, and 0.9, respectively. On the other hand, the corresponding results of $\{\mu_1, \mu_2, \mu_3\} = \{410, 450, 490\}$ are presented in Tables 4, 5 and 6.

The sample size calculations presented in Tables 1, 2, 3, 4, 5 and 6 reveal that, as expected, the computed sample sizes of the ANOVA procedure remain identical for different number of covariates P when all other factors are fixed. In contrast, the sample size of the exact approach

TABLE 1.
Computed sample size, estimated power, and simulated power for the ANOVA, approximate ANCOVA, and exact ANCOVA methods when $G = 3$, $\sigma_Y^2 = 10000$, $\rho = 0.1$, $\sigma^2 = 9900$, $\{\mu_1, \mu_2, \mu_3\} = \{400, 450, 500\}$, $\delta^2 = 0.1684$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$

P	ANOVA				Approximate ANCOVA				Exact ANCOVA			
	N_{TO}	Estimated power	Simulated power	Error	N_{TA}	Estimated power	Simulated power	Error	N_{TE}	Estimated power	Simulated power	Error
1	63	0.8148	0.8169	− 0.0021	63	0.8185	0.8158	0.0027	63	0.8115	0.8123	− 0.0008
2	63	0.8148	0.8062	0.0086	63	0.8181	0.7993	0.0188	63	0.8038	0.7995	0.0043
3	63	0.8148	0.8014	0.0134	63	0.8178	0.7922	0.0256	66	0.8176	0.8216	− 0.0040
4	63	0.8148	0.7913	0.0235	63	0.8174	0.7874	0.0300	66	0.8104	0.8094	0.0010
5	63	0.8148	0.7795	0.0353	63	0.8170	0.7796	0.0374	66	0.8029	0.8026	0.0003
6	63	0.8148	0.7703	0.0445	63	0.8166	0.7773	0.0393	69	0.8166	0.8185	− 0.0019
7	63	0.8148	0.7671	0.0477	63	0.8161	0.7622	0.0539	69	0.8093	0.8080	0.0013
8	63	0.8148	0.7516	0.0632	63	0.8157	0.7522	0.0635	69	0.8018	0.8038	− 0.0020
9	63	0.8148	0.7412	0.0736	63	0.8152	0.7476	0.0676	72	0.8156	0.8210	− 0.0054
10	63	0.8148	0.7355	0.0793	63	0.8147	0.7302	0.0845	72	0.8083	0.8049	0.0034

TABLE 2.

Computed sample size, estimated power, and simulated power for the ANOVA, approximate ANCOVA, and exact ANCOVA methods when $G = 3$, $\sigma_Y^2 = 10000$, $\rho = 0.5$, $\sigma^2 = 7500$, $\{\mu_1, \mu_2, \mu_3\} = \{400, 450, 500\}$, $\delta^2 = 0.2222$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$

P	ANOVA				Approximate ANCOVA				Exact ANCOVA			
	N_{TO}	Estimated power	Simulated power	Error	N_{TA}	Estimated power	Simulated power	Error	N_{TE}	Estimated power	Simulated power	Error
1	63	0.8148	0.9073	−0.0925	48	0.8136	0.7994	0.0142	48	0.8042	0.8006	0.0036
2	63	0.8148	0.9022	−0.0874	48	0.8130	0.7993	0.0137	51	0.8217	0.8215	0.0002
3	63	0.8148	0.9006	−0.0858	48	0.8123	0.7883	0.0240	51	0.8122	0.8062	0.0060
4	63	0.8148	0.8945	−0.0797	48	0.8115	0.7705	0.0410	51	0.8022	0.8053	−0.0031
5	63	0.8148	0.8926	−0.0778	48	0.8108	0.7592	0.0516	54	0.8201	0.8271	−0.0070
6	63	0.8148	0.8799	−0.0651	48	0.8099	0.7528	0.0571	54	0.8104	0.8177	−0.0073
7	63	0.8148	0.8737	−0.0589	48	0.8091	0.7366	0.0725	54	0.8004	0.7976	0.0028
8	63	0.8148	0.8653	−0.0505	48	0.8082	0.7196	0.0886	57	0.8185	0.8291	−0.0106
9	63	0.8148	0.8539	−0.0391	48	0.8072	0.7087	0.0985	57	0.8088	0.8157	−0.0069
10	63	0.8148	0.8526	−0.0378	48	0.8062	0.6957	0.1105	60	0.8263	0.8259	0.0004

TABLE 3.

Computed sample size, estimated power, and simulated power for the ANOVA, approximate ANCOVA, and exact ANCOVA methods when $G = 3$, $\sigma_Y^2 = 10000$, $\rho = 0.9$, $\sigma^2 = 1900$, $\{\mu_1, \mu_2, \mu_3\} = \{400, 450, 500\}$, $\delta^2 = 0.8772$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$

P	ANOVA				Approximate ANCOVA				Exact ANCOVA			
	N_{TO}	Estimated power	Simulated power	Error	N_{TA}	Estimated power	Simulated power	Error	N_{TE}	Estimated power	Simulated power	Error
1	63	0.8148	1.0000	−0.1852	15	0.8108	0.7696	0.0412	18	0.8751	0.8812	−0.0061
2	63	0.8148	1.0000	−0.1852	18	0.8934	0.8509	0.0425	18	0.8409	0.8485	−0.0076
3	63	0.8148	1.0000	−0.1852	18	0.8867	0.8026	0.0841	18	0.8007	0.8086	−0.0079
4	63	0.8148	1.0000	−0.1852	18	0.8784	0.7702	0.1082	21	0.8614	0.8628	−0.0014
5	63	0.8148	1.0000	−0.1852	18	0.8681	0.7050	0.1631	21	0.8272	0.8368	−0.0096
6	63	0.8148	1.0000	−0.1852	18	0.8548	0.6520	0.2028	24	0.8807	0.8873	−0.0066
7	63	0.8148	1.0000	−0.1852	18	0.8373	0.5948	0.2425	24	0.8516	0.8578	−0.0062
8	63	0.8148	1.0000	−0.1852	18	0.8136	0.5257	0.2879	24	0.8173	0.8263	−0.0090
9	63	0.8148	1.0000	−0.1852	21	0.9048	0.6459	0.2589	27	0.8735	0.8786	−0.0051
10	63	0.8148	1.0000	−0.1852	21	0.8904	0.5855	0.3049	27	0.8442	0.8505	−0.0063

TABLE 4.

Computed sample size, estimated power, and simulated power for the ANOVA, approximate ANCOVA, and exact ANCOVA methods when $G = 3$, $\sigma_Y^2 = 10000$, $\rho = 0.1$, $\sigma^2 = 9900$, $\{\mu_1, \mu_2, \mu_3\} = \{410, 450, 490\}$, $\delta^2 = 0.1077$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$

P	ANOVA				Approximate ANCOVA				Exact ANCOVA			
	N_{TO}	Estimated power	Simulated power	Error	N_{TA}	Estimated power	Simulated power	Error	N_{TE}	Estimated power	Simulated power	Error
1	96	0.8119	0.8143	− 0.0024	93	0.8023	0.7987	0.0036	96	0.8114	0.8156	− 0.0042
2	96	0.8119	0.8099	0.0020	93	0.8021	0.7925	0.0096	96	0.8063	0.8141	− 0.0078
3	96	0.8119	0.8016	0.0103	93	0.8019	0.7916	0.0103	96	0.8017	0.8075	− 0.0058
4	96	0.8119	0.7996	0.0123	93	0.8018	0.7863	0.0155	99	0.8109	0.8118	− 0.0009
5	96	0.8119	0.7929	0.0190	93	0.8016	0.7763	0.0253	99	0.8062	0.8090	− 0.0028
6	96	0.8119	0.7857	0.0262	93	0.8014	0.7802	0.0212	99	0.8014	0.7936	0.0078
7	96	0.8119	0.7836	0.0283	93	0.8012	0.7638	0.0374	102	0.8104	0.8122	− 0.0018
8	96	0.8119	0.7688	0.0431	93	0.8010	0.7641	0.0369	102	0.8057	0.8094	− 0.0037
9	96	0.8119	0.7729	0.0390	93	0.8008	0.7570	0.0438	102	0.8009	0.8004	0.0005
10	96	0.8119	0.7677	0.0442	96	0.8144	0.7638	0.0506	105	0.8099	0.8135	− 0.0036

TABLE 5.
Computed sample size, estimated power, and simulated power for the ANOVA, approximate ANCOVA, and exact ANCOVA methods when $G = 3$, $\sigma_Y^2 = 10000$, $\rho = 0.5$, $\sigma^2 = 7500$, $\{\mu_1, \mu_2, \mu_3\} = \{410, 450, 490\}$, $\delta^2 = 0.1422$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$

P	ANOVA				Approximate ANCOVA				Exact ANCOVA			
	N_{TO}	Estimated power	Simulated power	Error	N_{TA}	Estimated power	Simulated power	Error	N_{TE}	Estimated power	Simulated power	Error
1	96	0.8119	0.9125	−0.1006	72	0.8069	0.8032	0.0037	72	0.8007	0.7979	0.0028
2	96	0.8119	0.9072	−0.0953	72	0.8066	0.7901	0.0165	75	0.8122	0.8151	−0.0029
3	96	0.8119	0.9057	−0.0938	72	0.8063	0.7850	0.0213	75	0.8062	0.8072	−0.0010
4	96	0.8119	0.9057	−0.0938	72	0.8060	0.7835	0.0225	78	0.8177	0.8215	−0.0038
5	96	0.8119	0.8916	−0.0797	72	0.8057	0.7736	0.0321	78	0.8117	0.8011	0.0106
6	96	0.8119	0.8948	−0.0829	72	0.8054	0.7681	0.0373	78	0.8054	0.8039	0.0015
7	96	0.8119	0.8922	−0.0803	72	0.8051	0.7577	0.0474	81	0.8170	0.8215	−0.0045
8	96	0.8119	0.8818	−0.0699	72	0.8048	0.7561	0.0487	81	0.8109	0.8185	−0.0076
9	96	0.8119	0.8829	−0.0710	72	0.8044	0.7306	0.0738	81	0.8047	0.8041	0.0006
10	96	0.8119	0.8770	−0.0651	75	0.8218	0.7570	0.0628	84	0.8163	0.8204	−0.0041

TABLE 6.

Computed sample size, estimated power, and simulated power for the ANOVA, approximate ANCOVA, and exact ANCOVA methods when $G = 3$, $\sigma_Y^2 = 10000$, $\rho = 0.9$, $\sigma^2 = 1900$, $\{\mu_1, \mu_2, \mu_3\} = \{410, 450, 490\}$, $\delta^2 = 0.5614$, Type I error $\alpha = 0.05$, and nominal power $1 - \beta = 0.80$

P	ANOVA				Approximate ANCOVA				Exact ANCOVA			
	N_{TO}	Estimated power	Simulated power	Error	N_{TA}	Estimated power	Simulated power	Error	N_{TE}	Estimated power	Simulated power	Error
1	96	0.8119	1.0000	−0.1881	21	0.8084	0.7851	0.0233	24	0.8516	0.8516	0.0000
2	96	0.8119	1.0000	−0.1881	21	0.8035	0.7568	0.0467	24	0.8281	0.8296	−0.0015
3	96	0.8119	1.0000	−0.1881	24	0.8637	0.8023	0.0614	24	0.8022	0.8034	−0.0012
4	96	0.8119	1.0000	−0.1881	24	0.8600	0.7777	0.0823	27	0.8437	0.8500	−0.0063
5	96	0.8119	1.0000	−0.1881	24	0.8557	0.7548	0.1009	27	0.8203	0.8162	0.0041
6	96	0.8119	1.0000	−0.1881	24	0.8508	0.7160	0.1348	30	0.8586	0.8577	0.0009
7	96	0.8119	1.0000	−0.1881	24	0.8450	0.6800	0.1650	30	0.8375	0.8354	0.0021
8	96	0.8119	1.0000	−0.1881	24	0.8382	0.6346	0.2036	30	0.8141	0.8208	−0.0067
9	96	0.8119	1.0000	−0.1881	24	0.8301	0.6029	0.2272	33	0.8537	0.8593	−0.0056
10	96	0.8119	1.0000	−0.1881	24	0.8203	0.5493	0.2710	33	0.8325	0.8390	−0.0065

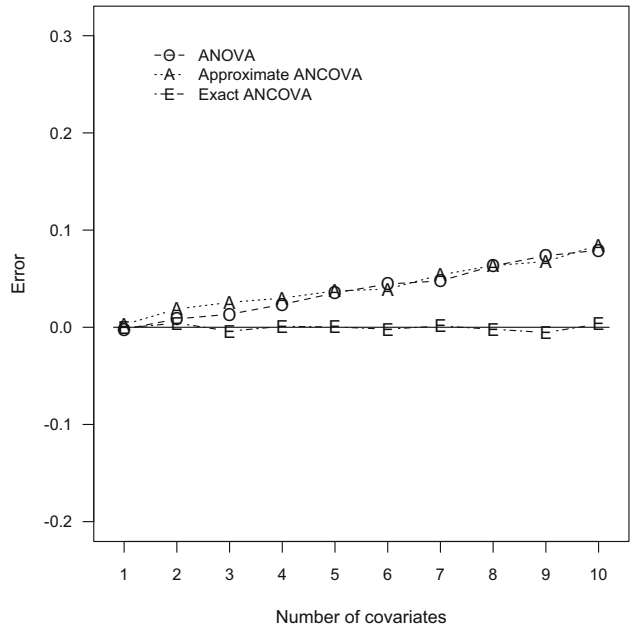

FIGURE 1.
Errors of power estimation for $G = 3$ and $\rho = 0.1$

increases with increase in the number of covariates P and with decrease in the effect size δ^2 when all other configurations are held constant. Likewise, the total sample size produced by the approximate procedure also increases with decrease in the effect size δ^2. However, the reported sample sizes in Tables 1 and 2 do not vary with P, and the computed sample sizes marginally increase with larger P for the other cases in Tables 3, 4, 5 and 6. More importantly, the total sample sizes N_{TO}, N_{TA}, and N_{TE} associated with the ANOVA, approximate ANCOVA, and exact ANCOVA procedures have a consistent order of $N_{TA} \leq N_{TO} \leq N_{TE}$ for all the cases in Tables 1 and 4 with $\rho = 0.1$. The order between the two sample sizes N_{TO} and N_{TE} is reversed for large magnitudes of $\rho = 0.5$ and 0.9 with $N_{TA} \leq N_{TE} \leq N_{TO}$ for the situations in Tables 2, 3, 5 and 6. For ease of explication, the estimated powers for the three different sample size procedures are also listed in Tables 1, 2, 3, 4, 5 and 6.

To justify the accuracy of sample size determination, Monte Carlo simulation studies were performed for the prescribed 60 design settings. With the computed sample sizes, parameter configurations, and nominal power, estimates of the true power were computed via Monte Carlo simulation of 10,000 independent data sets. For each replicate, N_{TO}, N_{TA}, and N_{TE} normal outcomes are generated with the ANCOVA models. Because the power function Ψ_E is irrelevant to the mean vector $\boldsymbol{\theta}$ and variance–covariance matrix $\boldsymbol{\Sigma}$ of the designated covariate distribution, the covariates are assumed to have independent and identical multinormal distribution $N_P(\mathbf{0}_P, \mathbf{I}_P)$ where $\mathbf{0}_P$ is a $P \times 1$ null column vector and $\mathbf{I}_P$ is an identity matrix of dimension P. The regression coefficients are chosen as $\beta_1 = \ldots = \beta_P = \beta^*$ and β^* is a designated value so that the resulting correlation $\rho = 0.1$, 0.5, and 0.9. Next, the Wald test statistic W^* was computed and the simulated power was the proportion of the 10,000 replicates whose test statistics W^* exceed the corresponding critical value $F_{2,\,\nu,\,0.05}$. The simulated power and error are also summarized in Tables 1, 2, 3, 4, 5 and 6 for all the ANCOVA designs. To illustrate the contrasting behavior of the three contending techniques, the induced errors for $\rho = 0.1$, 0.5, 0.9 in Tables 1, 2 and 3 are also plotted in Figs. 1, 2, and 3, respectively.

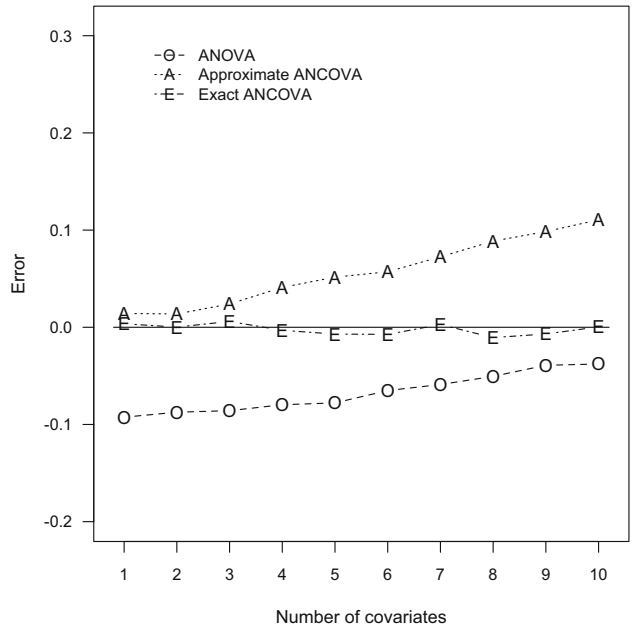

FIGURE 2.
Errors of power estimation for $G = 3$ and $\rho = 0.5$

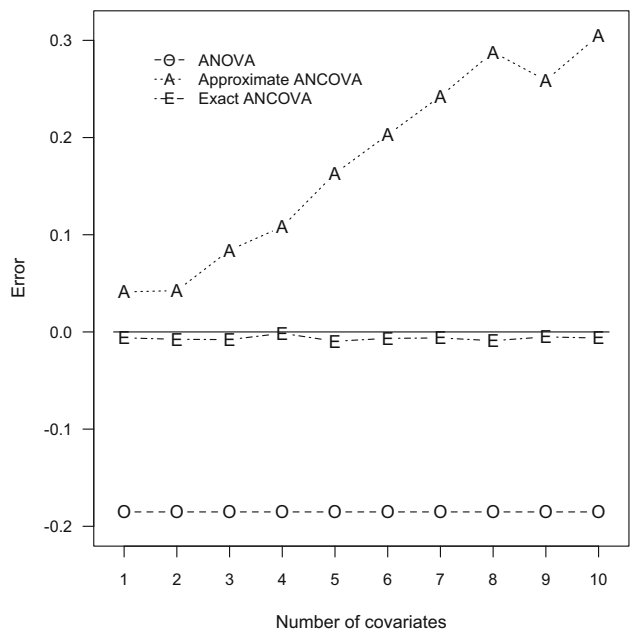

FIGURE 3.
Errors of power estimation for $G = 3$ and $\rho = 0.9$

According to the power comparisons, the ANOVA method generally does not provide accurate sample size calculations for an ANCOVA design. Unsurprisingly, the only exceptions occurred when the number of covariates is small and the correlation between the covariates and the response variable is close to zero as in Tables 1 and 4. The approximate ANCOVA method consistently gives larger power estimate than the simulated power for all cases considered here. The discrepancy noticeably increases with the number of covariates and the magnitude of effect size. The resulting errors can be as large as 0.0845, 0.1105, and 0.3049 associated with the scenarios of $P = 10$ in Tables 1, 2 and 3, respectively. For the relative smaller effect sizes in Tables 4, 5 and 6, the performance of the approximate ANCOVA formula has improved with the errors of 0.0506, 0.0628, and 0.2710 for the cases of $P = 10$. Consequently, the overestimation problem of the power function Ψ_A suggests that the computed sample sizes are generally inadequate to achieve the designated power level.

Regarding the accuracy of the proposed exact ANCOVA approach, the corresponding results in Tables 1, 2, 3, 4, 5 and 6 show that the differences between the estimated and simulated powers are fairly small. The largest absolute error is 0.0106 for the two cases of $P = 8$ and 5 in Tables 2 and 5, respectively. All the other 58 cases in Tables 1, 2, 3, 4, 5 and 6 have an absolute error less than 0.01. These numerical results imply that the proposed exact approach outperforms the ANOVA method and the approximate ANCOVA procedure for all design configurations considered here. Therefore, the suggested power and sample size calculations can be recommended for general use.

6. An Example

A documented example of Maxwell and Delaney (2004) is presented and extended next to demonstrate the usefulness of the suggested power and sample size procedures and accompanying software programs for the omnibus test of treatment effects in ANCOVA designs.

Specifically, Maxwell and Delaney (2004, Table 9.7, p. 429) provided the data for assessing the effectiveness of different interventions for depression. There are 10 participants with random assignment in each of the three intervention groups of (1) selective serotonin reuptake inhibitor (SSRI) antidepressant medication, (2) placebo, or (3) wait list control. The measurements are the pretest and posttest Beck Depression Inventory (BDI) scores of depressive individuals. The primary interest of the ANCOVA study is on the group differences of posttest BDI measurements using the pretest BDI scores as covariates. The results show that the estimates of adjusted group means and error variance are $\{\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\mu}_3^*\} = \{7.5366, 11.9849, 13.9785\}$ and $\hat{\sigma}^2 = 29.0898$, respectively. The omnibus F test statistic of treatment differences is $W^* = 3.73$, which yields a p-value of 0.0376. Therefore, the test result suggests that the intervention effects are significantly different at $\alpha = 0.05$. Although this is not the focus in the illustration of Maxwell and Delaney (2004), it can be computed from an ANOVA of posttest scores that the variance estimate is $\hat{\sigma}_Y^2 = 39.6185$. Hence, the sample squared correlation between the posttest and pretest BDI scores is $\hat{\rho}^2 = 1 - \hat{\sigma}^2\hat{\sigma}_Y^2 = 1 - 29.0898/39.6185 = 0.2658$. The observed value of the ANOVA F test of group differences is $F^* = 3.03$ with a p-value of 0.0647. At the significance level 0.05, the omnibus test of no intervention group difference on the posttest BDI scores cannot be rejected. Although null hypothesis significance testing is useful in various applications, it is important to consult the recent articles of Wasserstein and Lazar (2016) and Wasserstein et al. (2019) for the recommended principles underlying proper use and interpretation of statistical significance and p-values.

In view of the prospective nature of advance research planning, the general guidelines suggest published findings or expert opinions can offer reliable information for the vital characteristics of future study. Accordingly, it is prudent to adopt a minimal meaningful effect size in order

to enhance the generalizability of the result and the accumulation of scientific knowledge. For illustration, the prescribed summary statistics of the three-group depression intervention study are employed as population adjusted mean effects and variance component. The suggested power procedure shows that the resulting power for the omnibus test of group differences is $\Psi_E = 0.6145$ when the significance level α equals to 0.05. Because the computed power is substantially smaller than the common levels of 0.80 or 0.90, this implies that the group sample size $N = 10$ does not provide a decent chance of detecting the potential differences between treatment groups. To determine the proper sample size, the proposed sample size computations showed that the balanced group sample sizes of 15 and 19 are required to attain the nominal power of 0.8 and 0.9, respectively. The total sample sizes $N_T = 45$ and 57 are substantially larger than 30 of the exemplifying design. Essentially, it requires 50% and 90% increases of the sample size to meet the common power levels of 0.80 and 0.90, respectively. These design configurations are presented in the user specifications of the SAS/IML and R programs presented in the supplemental programs. Researchers can easily identify these statements and then modify the input values in the computer code to incorporate their own model characteristics.

7. Conclusions

ANCOVA provides a useful approach for combining the advantages of two widely established procedures of ANOVA and multiple linear regression. Despite the close resemblance among the three types of statistical analyses, their power computation and sample size determination are still theoretically distinct when the stochastic properties of the continuous covariates or predictors are taken into account. It is generally recognized that the use of ANCOVA may considerably reduce the number of subjects required than an ANOVA design to attain the required precision and power. For planning and evaluating randomized ANCOVA designs, an ANOVA-based sample size formula has been proposed in Cohen (1988) to accommodate the reduced error variance and degrees of freedom because of the use of effective and influential covariates. The procedure is very appealing from a computational standpoint and has been implemented in some statistical packages. However, no further analytical discussion and numerical evaluation are available to validate the appropriateness and implications of Cohen's (1988) method in the literature.

This article aims to address the potential limitation and approximate nature of the prevailing method and to describe an alternative and exact approach for power and sample size calculations in ANCOVA designs. It is demonstrated both theoretically and empirically that the seemly exact technique of Cohen (1988) does not involve all of the covariate properties in ANCOVA. Exact power and sample size procedures are described for the general linear hypothesis tests of treatment effects under the assumption that the covariate variables have a joint multinormal distribution. The simulation results reveal that the proposed technique is superior to the current method under a wide range of ANCOVA designs. More importantly, additional numerical assessments show that the suggested power function and sample size procedure preserve reasonably good performance under various non-normal situations, such as exponential, Gamma, Laplace, Log normal, uniform, and discrete uniform distributions. Hence, the proposed two-stage distribution and power function of the Wald statistic for the general linear hypothesis tests possess desirable robust properties and are also applicable to other continuous covariate distributions in various ANCOVA designs. Consequently, the presented methodology expands the power assessment and sample size determination of Shieh (2017) for contrast analysis in ANCOVA. To enhance the practical values, computer algorithms are also provided to facilitate the recommended power calculations and sample size determinations. With respect to the importance and implementation of random sampling, the fundamental and standard sampling designs and estimation methods can be found in Thompson (2012). Heterogeneity of variance is one of the unique and problematic factors known

as detrimental to the statistical inferences in ANCOVA (Harwell 2003; Rheinheimer and Penfield 2011). A potential topic for future study is to develop proper power and sample size procedures within the variance heterogeneity framework.

Acknowledgments

The author is grateful to the Associate Editor and referees for their valuable comments and suggestions which greatly improved the presentation and content of the article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Bratcher, T. L., Moran, M. A., & Zimmer, W. J. (1970). Tables of sample sizes in the analysis of variance. *Journal of Quality Technology*, *2*, 156–164. https://doi.org/10.1080/00224065.1970.11980429.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, *13*, 261–281. https://doi.org/10.2307/2527916.

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

Elashoff, J. D. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal*, *6*, 381–401. https://doi.org/10.2307/2527916.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Edinburgh: Oliver and Boyd.

Fleishman, A. I. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement*, *40*, 659–670. https://doi.org/10.1177/001316448004000309.

Fleiss, J. L. (2011). *Design and analysis of clinical experiments*. New York, NY: Wiley.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, *106*, 516–524. https://doi.org/10.1037/0033-2909.106.3.516.

Gupta, A. K., & Nagar, D. K. (1999). *Matrix variate distributions*. Boca Raton, FL: CRC.

Harwell, M. (2003). Summarizing Monte Carlo results in methodological research: The single-factor, fixed-effects ANCOVA case. *Journal of Educational and Behavioral Statistics*, *28*, 45–70. https://doi.org/10.3102/10769986028001045.

Huitema, B. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies* (Vol. 608). New York, NY: Wiley.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*, 422–431. https://doi.org/10.1037/0033-2909.99.3.422.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350–386. https://doi.org/10.3102/00346543068003350.

Kraemer, H. C., & Blasey, C. (2015). *How many subjects? Statistical power analysis in research* (2nd ed.). Los Angeles, CA: Sage.

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging*, *12*, 84–106. https://doi.org/10.1037//0882-7974.12.1.84.

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

Mendoza, J. L., & Stafford, K. L. (2001). Confidence interval, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement*, *61*, 650–667. https://doi.org/10.1177/00131640121971419.

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). New York, NY: Routledge.
Pavur, R., & Nath, R. (1984). Exact F tests in an ANOVA procedure for dependent observations. *Multivariate Behavioral Research*, *19*(408–420), 3. https://doi.org/10.1207/s15327906mbr1904_.
Pearson, E. S., & Hartley, H. O. (1951). Charts of the power function for analysis of variance tests, derived from the non-central F-distribution. *Biometrika*, *38*, 112–130. https://doi.org/10.2307/2332321.
Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology*, *34*, 383–392. https://doi.org/10.1037//0022-0167.34.4.383.
R Development Core Team. (2017). R: A language and environment for statistical computing [Computer software and manual]. Retrieved fromhttp://www.r-project.org.
Rheinheimer, D. C., & Penfield, D. A. (2011). The effects of Type I error rate and power of the ANCOVA F test and selected alternatives under nonnormality and variance heterogeneity. *Journal of Experimental Education*, *69*, 373–391. https://doi.org/10.1080/00220970109599493.
Rutherford, A. (2011). *ANOVA and ANCOVA: A GLM approach*. Hoboken, NJ: Wiley.
Ryan, T. P. (2013). *Sample size determination and power*. Hoboken, NJ: Wiley.
Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, *69*, 682–689. https://doi.org/10.2307/2286002.
SAS Institute. (2017). *SAS/IML user's guide 14.3*. Cary, NC: SAS Institute Inc.
Scariano, S. M., & Davenport, J. M. (1987). The effects of violations of independence assumptions in the one-way ANOVA. *The American Statistician*, *41*, 123–129. https://doi.org/10.2307/2684223.
Scheffe, H. (1961). *The analysis of variance*. New York, NY: Wiley.
Shieh, G. (2006). Exact interval estimation, power calculation and sample size determination in normal correlation analysis. *Psychometrika*, *71*, 529–540. https://doi.org/10.1007/s11336-04-1221-6.
Shieh, G. (2007). A unified approach to power calculation and sample size determination for random regression models. *Psychometrika*, *72*, 347–360. https://doi.org/10.1007/s11336-007-9012-5.
Shieh, G. (2017). Power and sample size calculations for contrast analysis in ANCOVA. *Multivariate Behavioral Research*, *52*, 1–11. https://doi.org/10.1080/00273171.2016.1219841.
Thompson, S. K. (2012). *Sampling*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118162934.
Tiku, M. L. (1967). Tables of the power of the F-test. *Journal of the American Statistical Association*, *62*, 525–539. https://doi.org/10.2307/2283980.
Tiku, M. L. (1972). More tables of the power of the F-test. *Journal of the American Statistical Association*, *67*, 709–710. https://doi.org/10.2307/2284473.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133. https://doi.org/10.1080/00031305.2016.1154108.
Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p<0.05$". *The American Statistician*, *73*, 1–19. https://doi.org/10.1080/00031305.2019.1583913.
Yang, H., Sackett, P. R., & Arvey, R. D. (1996). Statistical power and cost in training evaluation: Some new considerations. *Personnel Psychology*, *49*, 651–668. https://doi.org/10.1111/j.1744-6570.1996.tb01588.x.

Manuscript Received: 19 APR 2019
Final Version Received: 19 NOV 2019