Behav Genet (2017) 47:255–261
DOI 10.1007/s10519-016-9828-9

BRIEF COMMUNICATION

A Power Calculator for the Classical Twin Design

Brad Verhulst[1]

Received: 19 August 2016 / Accepted: 3 November 2016 / Published online: 19 November 2016

Abstract Power is a ubiquitous, though often overlooked, component of any statistical analyses. Almost every funding agency and institutional review board requires that some sort of power analysis is conducted prior to data collection. While there are several excellent on line power calculators for independent observations, twin studies pose unique challenges that are not incorporated into these algorithms. The goal of the current manuscript is to outline a general method for calculating power in twin studies, and to provide functions to allow researchers to easily conduct power analyses for a range of common twin models. Several scenarios are discussed to demonstrate the importance of various factors that influence the power within the classical twin design and to serve as examples for the provided functions.

Keywords Power · Biometrical genetics · Twin study · Variance components

Introduction

Power is the probability of rejecting the null hypothesis when the null hypothesis is false (Cohen 1988). Accordingly, power is an essential component of any statistical analysis. While there are a number of high quality online power calculators for a variety of linear models, power analyses in twin models can be more complicated than other types of linear models and existing methods often do not adequately capture the additional complexity of the twin context. Existing presentations of statistical power for twin and family studies present a variety of guidelines and power tables (Posthuma and Boomsma 2000; Neale et al. 1994; Martin et al. 1978), but do not allow the reader to rapidly check the power of a specific set of parameters in a univariate and bivariate twin twin study. One exception to this comes from Visscher (2004), where an online power calculator allows users to quickly test power for continuous, univariate twin models. The power to detect a specific variance component (e.g. the additive genetic variance component) in a twin study depends upon the other variance components in the model (e.g. the common and unique environmental variance components). To address this I have prepared series of functions to conduct power analyses for a variety of common twin models. The functions use the R statistical environment (R Development Core Team 2008), and OpenMx in particular (Neale et al. 2015; Boker et al. 2015, 2011).

In the sections that follow, I present the theory and algebra used calculate power in the classical twin design for the univariate and bivariate cases. I then discuss several typical scenarios where the functions can be used to conduct a power analysis that are common in the twin literature.

The framework to conduct power analyses in twin models

Figure 1 presents a graphical depiction of primary components of statistical power: α or the probability of a Type I Error; β or the probability of a Type II Error, N or sample size; and d or effect size. Figure 1a presents the standard figure typically used to discuss statistical power with a

Edited by John K Hewitt.

✉ Brad Verhulst
brad.verhulst@gmail.com

[1] Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA

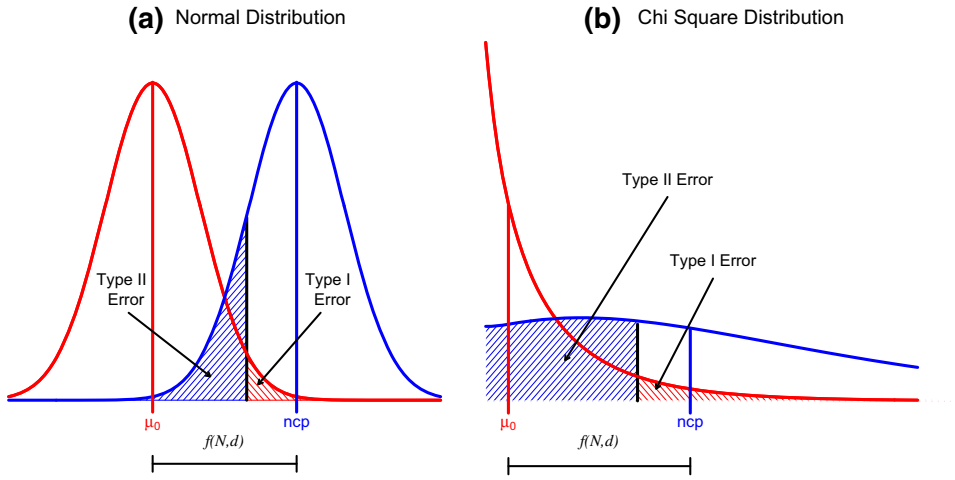

Fig. 1 A graphical depiction of primary components of statistical power. The *red* distribution represents the distribution of test statistics under the null and the *blue* distribution represents the distribution of test statistics under the alternative. The *black line* indicates the significance threshold. The hatched *red lines* indicate the probably of a Type I Error or α Error. The hatched *blue lines* indicate the probably a Type II Error or β Error. Power is $1-\beta$. The non-centrality parameter (ncp), is the mean of the test statistic distribution under the alternative (Color figure online)

normal distribution and Figure 1b presents the analogous figure for a chi-square (χ^2) distribution. Because Likelihood ratio tests (LRTs) are the primary method of hypothesis testing for twin analyses, and the LRTs rely on the χ^2 distribution, it is necessary to discuss power from the perspective of the χ^2 distribution. The red (left or taller in Fig. 1a, b, respectively) distribution represents the distribution of test statistics assuming the null hypothesis is true while the blue (right or flatter) distribution represents the distribution of test statistics assuming the alternative hypothesis is true. For both panels in Fig. 1 the black line indicates the significance threshold, the hatched red lines to the right of the significance threshold indicate the probably of rejecting the null hypothesis when the null hypothesis is true (a Type I Error), and the hatched blue lines to the left of the significance threshold indicate the probably of failing to reject the null hypothesis, when the alternative hypothesis is true (a Type II Error). Power is the area under the alternative (blue) distribution that is not hatched. When the normal (or similar) distribution is used for hypothesis testing, the null and alternative distributions look quite similar. For the χ^2 distribution, however, the distribution under the null looks substantially different from the distribution under they alternative. Importantly, the components of statistical power work the same way in both scenarios.

Standard power calculations set the desired level of power at .80 and the Type I Error rate at $\alpha = .05$. In twin models in particular, but in any model where there is a strong theoretical boundary on a parameter, the Type I Error rate will be over estimated. Because variances must be positive, the Type I Error rate for a single variance component is actually a mixture of a χ^2 distribution with 1 df and a χ^2 distribution with 0 df (Wu and Neale 2012; Visscher 2006; Dominicus et al. 2006). It turns out that for the 1 df case, the solution for the mixture reduces to setting $\alpha = .10$. For the general multiple df tests, there is no straightforward correction for the Type I Error rate, and therefore must be calculated empirically. Dominicus et al. (2006) provide an analytical solution for the specific case of comparing the full ACE with the E only model (2 df test). Thus, not taking the mixture of χ^2 distributions into consideration with multiple df tests will lead to an under estimate power.

The current discussion of power focuses on the non-central χ^2 distribution. The mean of the non-central χ^2 distribution, or the non-centrality parameter (ncp), is the sum of the mean of the test statistic distribution under the alternative hypothesis and the degrees of freedom. Two features of the ncp that are integral to the current discussion. First, as the effect size gets larger, the mean of the test statistic gets larger, and the ncp gets larger. Second, as sample size increases, the standard deviation of the sampling distribution of the test statistic gets tighter, and the ncp gets larger. Thus, power will increase with both larger effect sizes and larger sample sizes.

The general procedure for the twin power analysis has 4 steps. The first step is to simulate twin data that corresponds with the expected results. These expectations should be based, as far possible, on the literature. While users will not have to simulate the data themselves, they will need to provide the proportions of variance for the standardized A, C and E variance components (the function

will automatically simulate data based on these values). At this stage it is important to consider the ratio of MZ to DZ twins, as the power to detect significant genetic or environmental variance is influenced by this ratio. See Visscher (2004) for a more detailed discussion of the ratio between MZ and DZ twins and power.

The second step is to fit the full and reduced models to the simulated data to obtain the χ^2 value from the likelihood ratio test. As a check, the fitted parameter estimates are returned. It is important to make sure that the fitted parameters correspond to the values that you simulated in case there was some problem with estimation. When extreme values are chosen for the variance components, small sample sizes are specified, or complex models are utilized, the fitted parameters may not correspond with the specified simulated values. In these cases, it is possible to increase the sample size (keeping the ratio of the MZ to DZ twins constant). This will not affect the estimates of the required sample size or the power.

The third step is to calculate the weighted non-centrality parameter (Wncp). To do this, we divide the χ^2 value by the total sample size ($N_M Z + N_D Z$). By dividing the χ^2 value by the sample size, we are calculating the average contribution of each family to χ^2. Therefore, this value is dubbed the weighted ncp.

The final step is to calculate power. The essential component to discussing power from the perspective of the ncp is that the ncp increases linearly with sample size. For example, if the value of the χ^2 value for the LRT between an ACE model and an AE model is 10 with 500 MZ and 500 DZ twin pairs ($N = 1000$), each family will contribute $10/1000 = .01$ to the χ^2 value, on average. It is then possible to extrapolate that with 2000 families, the χ^2 value would be 20, and with 500 families the χ^2 value would be 5. Therefore, we can multiply the Wncp obtained in the previous step by a vector of sample sizes to obtain a vector of ncp's. This vector of ncp's is then used to calculate the power for a range of sample sizes. The power can then be plotted to obtain a standard power graph or the required sample size for a specific level of power can be obtained.

Because the family is the unit of measurement for twin studies, the sample size here refers to the total number of families (Nmz + Ndz) rather than the number of individual twins. To obtain the number of MZ (or DZ) twins, you must multiply the total sample size by the proportion of the sample that was specified to be MZ (or DZ) twins.

Demonstrations

To demonstrate the application of the power analysis functions, power analyses for five common scenarios were conducted. These examples are not intended to be exhaustive, but instead highlight a few considerations that influence power twin studies. A complete description of the functions used to conduct the power analysis and a tutorial can be found at: http://www.people.vcu.edu/~bverhulst/power/power.html.

The first demonstration examines the power to detect a moderate sized standardized A (or C) variance component when the magnitude of the complimentary C (or A) variance component was varied. Specifically, the common environmental (or genetic) variance was tested at .1, .3, or .5 proportion of the phenotypic variance for small, medium, and large effect sizes, respectively. Figure 2 presents the results of the power analysis. As can be seen, the power to detect both variance components depend on the magnitude of the other variance component. As the opposing variance component increases, the power to detect the variance component of interest increases. Interestingly, the increase in power is not symmetrical across A and C. The power increase is larger for the A as C increases, than it is for C as A increases. Therefore, when conducting power

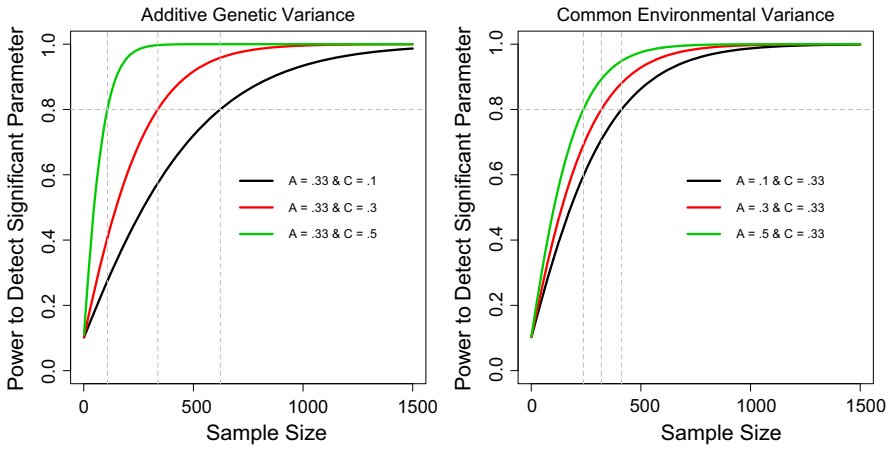

Fig. 2 Power analysis for the additive genetic and common environmental variance components as the level of the other variance component varies. In the left panel, the A variance component is set to .33 and the C variance component varies from .1 to .3 to .5. In the right panel, C is set to .33 and A varies from .1 to .3 to .5. The sample size for MZ and DZ twins was equal (Color figure online)

analyses for twin studies, it is essential to consider not only the level of A but also the level of C.

The second demonstration examines the power to detect the A and C variance components when the ratio of MZ to DZ sample size varied. Specifically, the ratio of MZ to DZ twins varied from 5:1 to 1:1 to 1:5. While these ratios are extreme, they clearly illustrate the impact of differential MZ to DZ sample size ratios. As can be seen in the left panel of Fig. 3, the power to detect A is maximized when there are approximately equal numbers of MZ and DZ twins. Deviations from a 1:1 ratio in either the MZ or DZ direction, reduce the power to detect a significant A component. By contrast, the power to detect C is highest if there are a surplus of DZ relative to MZ twins, but the increase in power is minimally better than an equal MZ:DZ ratio. A surplus of MZ twins is strongly reduces the power to detect C. This highlights the importance of considering the ratio of MZ to DZ twins when conducting a twin power analysis. For a more complete discussion of the optimal ratio of MZ to DZ twins, see Visscher (2004).

The third power analysis demonstration examines the power to detect a significant A variance component using continuous data relative to binary data with prevalences ranging from .5 to .05. For this example we kept the A and C variance components at .33 and used equal numbers of MZ and DZ twins. As can be seen in Fig. 4, there is a large reduction in power for a median split (prevalence = .50) relative to a continuous variable, and as the prevalence of a phenotype decreases, the power to detect A decreases.

The fourth example of a power analysis examines the power to detect a significant genetic correlation (Rg) between two phenotypes. As the power to detect a significant Rg depends on the magnitude of the genetic variance in the phenotypes under examination, the values of A, which were equated across phenotypes, were varied from .3 to .4 to .5 and the values of Rg were varied from .1 to .3 to .5. The value of C for both phenotypes was set to .33 and the value of E was adjusted to ensure that the variance of all of the traits was 1. The left column of Fig. 5 presents the power curves for the A variance component in the first phenotype. The right column presents the power curves for Rg. As can be seen, the power to detect significant Rg increases as the magnitude of Rg increases and as the magnitude of A increases.

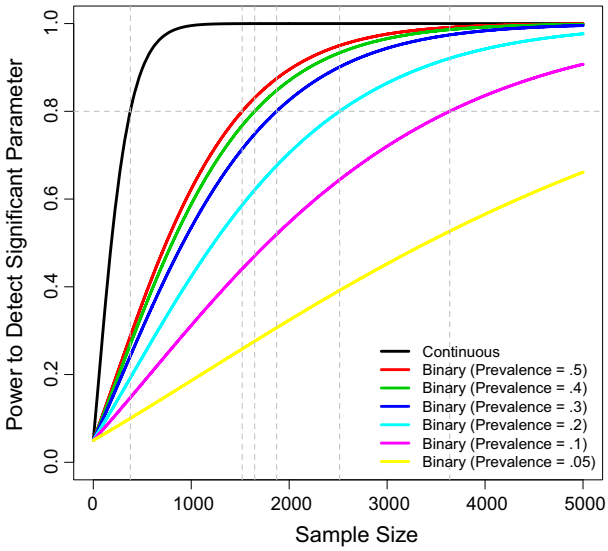

Fig. 4 The power to detect a significant A variance component for continuous data relative to binary data with prevalences decreasing from .5 to .05. The values of A and C are equal and the sample size for MZ and DZ twins was equal (Color figure online)

There are many common bivariate models that could be examined, and multiple potentially interesting parameters within each model. For example, the power to detect genetic variance in one phenotype if it is (1) genetically correlated with another phenotype and (2) the genetic variance of the second phenotype varies. Under such a scenario, it is important to remember that logically, if the genetic variance in the second phenotype goes to zero, the genetic correlation necessarily goes to zero as well.

Finally, the fifth demonstration examines the power to detect significant differences between the variance components for males and females, often called sex limitation

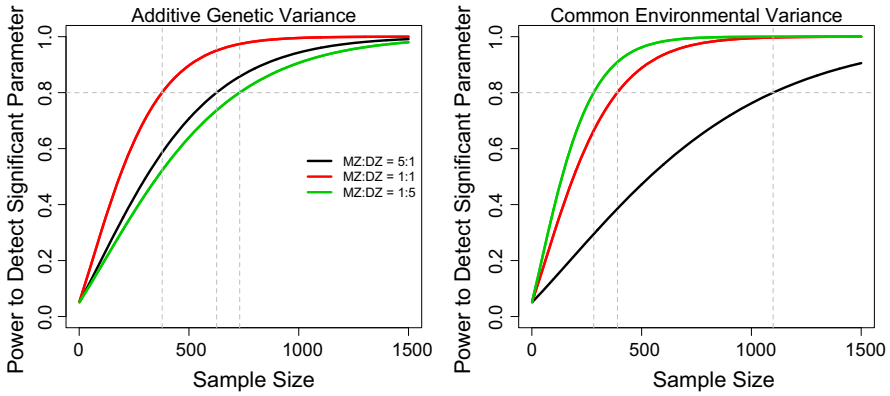

Fig. 3 Power analysis for the additive genetic and common environmental variance components as a function of the ratio of MZ to DZ twins. The power to detect A is presented in the *left* panel and the power to detect C is presented in the *right* panel. The values of A and C are equal (Color figure online)

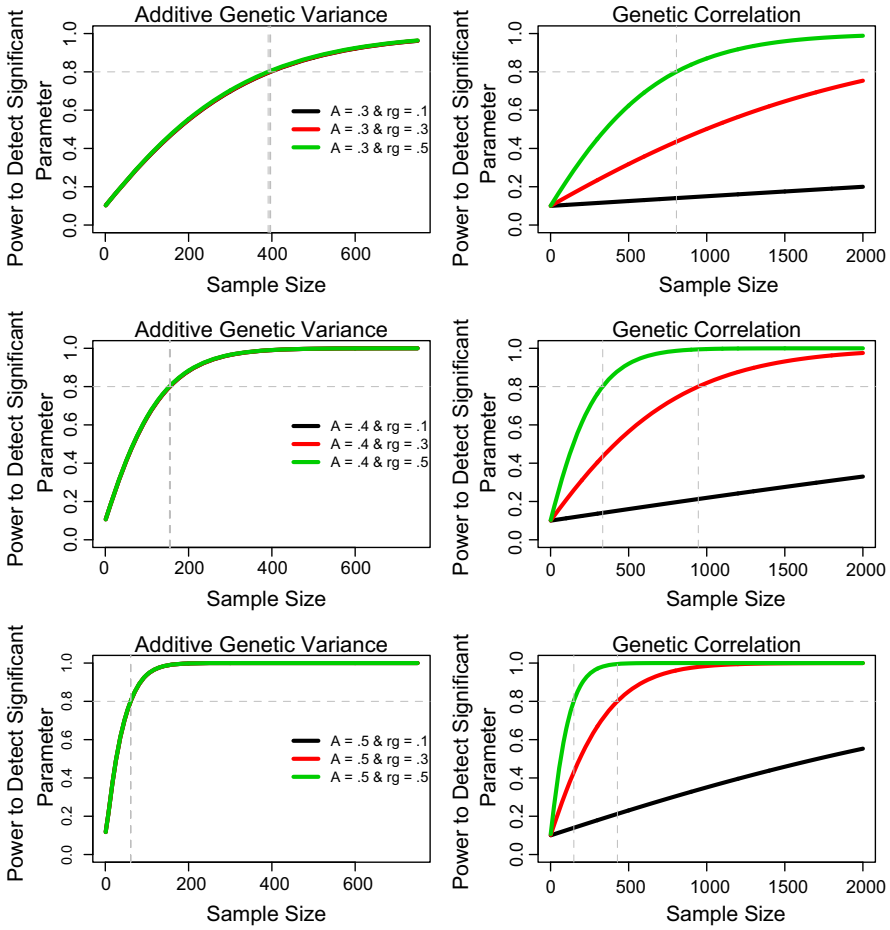

Fig. 5 Power to detect a significant genetic correlation between two phenotypes. The values of A for both phenotypes are are fixed at .3, .4 and .5 in the *top*, *middle* and *bottom* rows, respective. The power to detect a significant A variance component for the first phenotype is presented in the *left* column. As the phenotypes had the same values of A, the power to detect the A in the second phenotype was equivalent to the first. The Rg between the phenotypes was varied from .1, to .3 to .5. The value of C was fixed at .33 for both phenotypes in all conditions and the Rc (correlation between the common environment) mirrored the Rg. The variance of E was computed so that the total variance of each trait summed to 1. There were no correlations between the unique environment components for the phenotypes (Re = 0). The sample size for MZ and DZ twins was equal (Color figure online)

models (Medland 2004; Neale et al. 2006; Harris 1948). There are two distinct forms of sex limitation that are commonly discussed in the literature. The first type of sex limitation, qualitative sex limitation, assesses the extent to which the same genetic factors contribute to phenotypic variation in both sexes. The other type of sex limitation, quantitative sex limitation, assesses the same genetic factors contribute to differing amounts of phenotypic variance in each sex. In both cases, larger proportions of genetic variation will increase the power to detect sex limitation.

The demonstration of qualitative sex limitation is presented in the left panel of Fig. 6. For the demonstration, the proportion of genetic variance was set at .5 and the proportion of shared environmental variance was set at .1 for both males and females. The proportion of shared genetic variance (or Rg_{mf}) between males and females tested at .9, .7, .5, .3 and .1. As can be seen in the figure, there is very little power to detect qualitative sex limitation when the correlation between the proportion of shared genetic variance between males and females is high, but increases to reasonable levels when only a few of the genetic factors that contribute to the phenotype in males also contribute to the phenotype in females.

The demonstration of quantitative sex limitation is presented in the right panel of Fig. 6. For this demonstration, the proportion of common environmental variation is small and equal for both sexes ($C_m = C_f = .2$), the proportion of additive genetic variation in females is moderate ($A_f = .5$), and the proportion of additive genetic variation in males varies from $A_m = .2$ to $A_m = .3$ to $A_m = .4$. As can be seen in the figure, as additive genetic estimates for males and females diverge, the power to detect differences in the additive genetic estimates for each sex. This type of analysis can also be done for common and unique environmental quantitative sex differences. It is important to note that for these values for both qualitative and quantitative sex limitation, over 1000 twin pairs is necessary to detect genetic sex limitation with a reasonable magnitude.

Discussion

The preceding sections present a framework for conducting power analyses in twin models. Five examples are discussed that highlight some relevant considerations. Specifically, the power to detect a given level of A depends

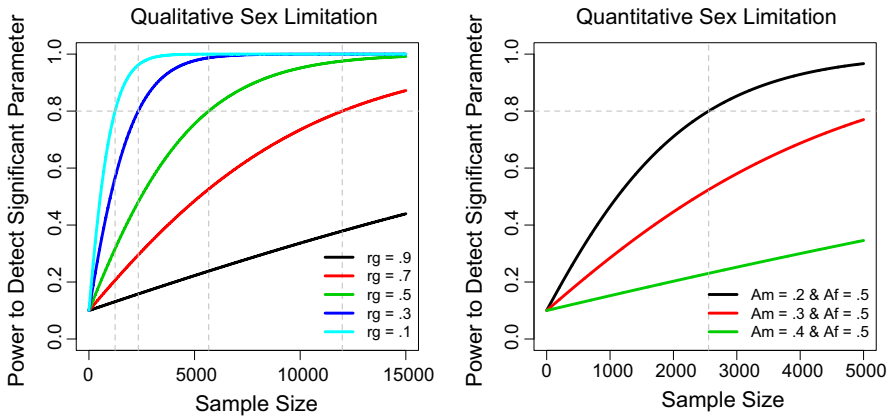

Fig. 6 Power to detect a significant qualitative and quantitative sex limitation. In the power analysis demonstration for qualitative sex limitation presented in the *left* panel, the proportion of genetic variance was set at .5 and the proportion of shared environmental variance was set at .2 for both males and females. The proportion of shared genetic variance between males and females was varied from .8 to .6 to .4. In the demonstration for quantitative sex limitation presented in the *right* panel, the proportion of shared environmental variance was fixed at .2 for both sexes, the proportion of genetic variance in females was fixed at .5, and the proportion of genetic variance in males varied from .2 to .3 to .4. The proportion of variance for E was computed so that the total variance of each trait summed to 1 in all cases.eps (Color figure online)

on the assumed value of C, and vice versa. The power to detect A and C depend on the ratio of MZ to DZ twins, with approximately equal proportions of MZ and DZ twins providing optimum levels of power. There is more power to detect variance components from continuous variables relative to binary variables, and as the prevalence of a binary phenotype decreases, so too does the power to detect the variance components. The power to detect significant genetic correlations depends on the magnitude of the additive genetic components of each constituent phenotype. Finally, the power to detect genetic sex limitation is fairly low and may require a substantial number of families. While these examples present common considerations, the potential permutations of these scenarios are virtually infinite and the data available to researchers is often quite specific. Accordingly, most people will prefer to conduct a limited number of power analyses that reflect their specific data. The functions and scripts used to conduct these power analyses are available at: http://www.people.vcu.edu/~bverhulst/power/power.html.

Notably, the power analyses presented here intentionally do not cover multivariate genetic models, such as the Cholesky decomposition, the independent or the common pathway models. While these models are common in the literature, calculating power for such models is not straight forward because determining the Type I Error rate is ambiguous for LRTs with multiple degrees of freedom if there are theoretical boundaries for the parameter estimates. Specifically, for most multivariate hypothesis tests the Type I Error rate must be empirically estimated from a mixture of multiple χ^2 distributions with different degrees of freedom. Thus, the Type I Error rate for a multivariate genetic models with 7 df is a complex mixture of 8 χ^2 distributions with degrees of freedom ranging sequentially from 0 to 7.

To accommodate individuals interested in conducting power analyses for multivariate models, functions to simulate data for the common application of the Cholesky decomposition, independent and common pathway models are provided. Users can then follow the steps delineated above and insert the data into the appropriate multivariate scripts to calculate the difference in the log likelihood between the saturated and reduced models of interest, and divide that χ^2 by N to obtain the Weighted ncp. An example of how to conduct a power analysis similar to this is discussed in the on line tutorial. Those interested in conducting such power analyses should do so with caution and at their own peril.

The discussion to this point has focused on *a priori* power analyses (or power analyses conducted before a grant is submitted and data is collected). Another common usage of power analyses is *post hoc* power analysis, where the values obtained from a specific sample are used to calculate the power to detect a significant effect. To conduct a *post hoc* power analysis it is possible to insert the obtained χ^2 values and sample sizes from a completed analysis into the functions provided. Specifically, the difference in the likelihood for the full and the reduced models (as estimated in the data), can be divided by the observed sample size to obtain the weighted ncp, in the same way as was described above. This weighted ncp can be used to calculate the power for a range of sample sizes.

The functions used to conduct these power analysis and a tutorial can be found at: http://www.people.vcu.edu/~bverhulst/power/power.html.

Supporting information

Power scripts

The functions used to fit all of the examples described in the current paper available on line at http://www.people.vcu.edu/~bverhulst/power/power.html.

Acknowledgments An earlier version of this paper was presented at the 2016 International Twin Workshop, March 10th, 2016. The author would like to thank the workshop faculty and students for their suggestions to improve the paper. This research was supported by by R25MH-019918 (PI: Hewitt), R01DA-018673 (PI: Neale) and R25DA-26119 (PI: Neale).

Compliance with ethical standards

Conflict of interest Brad Verhulst declares that he has no conflicts of interest.

Human and animal rights and informed consent This article does not contain any studies with human or animal participants performed by any of the authors.

References

Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T, Fox J (2011) Openmx: an open source extended structural equation modeling framework. Psychometrika 76(2):306–317

Boker SM, Neale MC, Maes HH, Wilde MJ, Spiegel M, Brick TR, Driver C (2015) Openmx 2.3.1 user guide [Computer software manual]

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Mahwah

Dominicus A, Skrondal A, Gjessing HK, Pedersen NL, Palmgren J (2006) Likelihood ratio tests in behavioral genetics: problems and solutions. Behav Genet 36(2):331–340. doi:10.1007/s10519-005-9034-7

Harris H (1948) On sex limitation in human genetics. Eugen Rev 40(2):70–76

Martin NG, Eaves LJ, Kearsey MJ, Davies P (1978) The power of the classical twin study. Heredity 40(1):97116

Medland SE (2004) Alternate parameterization for scalar and non-scalar sex-limitation models in Mx. Twin Res 7(3):299–305

Neale MC, Eaves LJ, Kendler KS (1994) The power of the classical twin method to resolve variation in threshold traits. Behav Genet 24:239–258

Neale MC, Hunter MD, Pritikin JN, Zahery M, Brick TR, Kickpatrick RM, Boker SM (2015). OpenMx 2.0: extended structural equation and statistical modeling. *Psychometrika*. doi: 10.1007/s11336-014-9435-8

Neale MC, Rysamb E, Jacobson K (2006) Multivariate genetic analysis of sex limitation and g x e interaction. Twin Res Hum Genet 9(4):481–489. doi:10.1375/183242706778024937

Posthuma D, Boomsma DI (2000) A note on the statistical power in extended twin designs. Behav Genet 30(2):147–158

R Development Core Team (2008) R: A language and environment for statistical computing [Computer software manual]. Vienna. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Visscher PM (2004) Power of the classical twin design revisited. Twin Res 7(5):505–512

Visscher PM (2006) A note on the asymptotic distribution of likelihood ratio tests to test variance components. Twin Res Hum Genet 9(4):490–495. doi:10.1375/183242706778024928

Wu H, Neale MC (2012) Adjusted confidence intervals for a bounded parameter. Behav Genet 42(6):886–898. doi:10.1007/s10519-012-9560-z