Twin Research and Human Genetics (2020), **23**, 87–89
doi:10.1017/thg.2020.46

CAMBRIDGE UNIVERSITY PRESS

Article

Statistical Power and the Classical Twin Design

Pak C. Sham[1], Shaun M. Purcell[2], Stacey S. Cherny[3,4], Michael C. Neale[5,6] and Benjamin M. Neale[7,8,9]

[1]Centre for PanorOmic Sciences, State Key Laboratory of Brain and Cognitive Sciences, and Department of Psychiatry, LKS Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong, [2]Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [3]Department of Epidemiology and Preventive Medicine, Tel Aviv University, Tel Aviv-Yafo, Israel, [4]Department of Psychiatry, LKS Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong, [5]Department of Biological Psychology, Vrije Universiteit, Amsterdam, the Netherlands, [6]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA, [7]Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, [8]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA and [9]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Abstract

Dr Nick Martin has made enormous contributions to the field of behavior genetics over the past 50 years. Of his many seminal papers that have had a profound impact, we focus on his early work on the power of twin studies. He was among the first to recognize the importance of sample size calculation before conducting a study to ensure sufficient power to detect the effects of interest. The elegant approach he developed, based on the noncentral chi-squared distribution, has been adopted by subsequent researchers for other genetic study designs, and today remains a standard tool for power calculations in structural equation modeling and other areas of statistical analysis. The present brief article discusses the main aspects of his seminal paper, and how it led to subsequent developments, by him and others, as the field of behavior genetics evolved into the present era.

Keywords: Twin studies; statistical power; type I error; type II error; null hypothesis; research design; statistical significance; effect size

(Received 24 February 2020; accepted 16 April 2020)

Dr Nick Martin is one of the most prolific and influential behavioral geneticists in the world, who has also been a key motivator, teacher and role model for his students, including ourselves. Over the years, we have greatly benefitted from Nick's wonderful teaching, very often demonstrating how theory can be applied in practice to investigate interesting and important scientific questions, and providing a much-needed historical perspective on the latest developments in our fast-moving field. It is therefore our great honor and privilege to review one of Nick's earliest papers, in celebration of his 70th birthday.

The paper, 'The power of the classical twin study' (Martin et al., 1978), was based on work from Nick's PhD thesis (Martin, 1976), completed in the Department of Genetics at the University of Birmingham. It was in this department that the field of biometrical genetics (Evans et al., 2002; Mather & Jinks, 1982) was established by pioneers who included Kenneth Mather, John Jinks, David Fulker and Lindon Eaves (Nick's PhD supervisor). The principles of biometrical genetics, as compared to other contemporary approaches to the analysis of family data, were laid down in a seminal paper from that department (Jinks & Fulker, 1970).

While the aim of biometrical genetics was to partition the sources of individual differences in the population according to various genetic and environmental sources of variation, Jinks and Fulker recognized that the ability to untangle different sources of variation from one another requires certain minimal experimental conditions — the 'minimum data.' For example, an analysis of variance for monozygotic (MZ) twins reared apart would yield two summary statistics, the between-group and the within-group mean-squares, which, when equated to the theoretical expected mean squares under the classical quantitative genetic model, would provide estimates for the total genetic and the total environmental variances. However, this analysis would not be able to separate out additive genetic effects from those of genetic dominance, nor could it distinguish the familial environment shared by siblings reared together from environmental influences unique to each sibling. A study that included a wider variety of relationships would provide more summary statistics, which would enable more sources of variation to be jointly estimated from the data.

Martin et al. (1978) recognized that even when an experimental design could provide the 'minimum data' for resolving certain sources of variation, the probability of achieving this in practice would still depend on having a sufficient sample size. To quote, 'If the power of a study to detect a given effect is low and in fact we do not find evidence for the effect in our sample then we should be foolish to infer that the effect is not present in the population' (p. 99). This remark is equivalent to the ever-valid saying 'Absence of evidence is not evidence of absence.' They pointed out that theoretical power calculations in the literature at the time dealt with 'human experimental designs which are seldom (if ever) used' but not 'the classical twin design, the most common design in human biometrical genetics' (p. 99).

The paper then went on to describe an analytical approach to perform a power calculation for the classical twin design. The method involved calculating the expected values of the observed

Author for correspondence: Pak C. Sham, Email: pcsham@hku.hk

Cite this article: Sham PC, Purcell SM, Cherny SS, Neale MC, and Neale BM. (2020) Statistical Power and the Classical Twin Design. *Twin Research and Human Genetics* **23**: 87–89, https://doi.org/10.1017/thg.2020.46

mean squares under the specified parameter values of a true model, and then equating these to the theoretical expected mean squares under a false model to estimate the parameters of the false model (using iterative weighted least squares). By substituting the expected mean squares under the true model as the observed mean squares of a goodness-of-fit chi-square test statistic for the false model, they obtained the noncentrality parameter of the (typically chi-squared) distribution of the test statistic. This enabled them to calculate the approximate power of the test for any desired significance level. Because the noncentrality parameter is proportional to sample size, the results can be easily extrapolated to calculate the power for any sample size, and to calculate the required sample size for any desired power. The accuracy of the power estimates obtained from the noncentral chi-squared distribution was shown to be acceptable by simulation for a range of parameter values and sample sizes. Using this method, it was shown that 600 twin pairs were required to reject most false models and that an optimal proportion of monozygotic (MZ) and dizygotic (DZ) twin pairs under most true models was between $\frac{1}{3}$ and $\frac{1}{2}$. The paper ended with a section on the power of detecting nonadditive and directional effects, with three subsections: (1) G × E interaction, by regressing pair variances on pair means, (2) directional dominance, by testing the phenotypic distribution for skewness and (3) directional allele frequency differences, again by testing the phenotypic distribution for skewness. The scoring of many behavioral and psychological tests often results in non-normal distributions of sum or factor scores, which can bias all three of these tests, but they still have potential for other variables such as neuroimaging measures, whose distributions accord better with the central limit theorem.

Two other papers from Nick and colleagues published at around the same time (Eaves et al., 1978; Martin & Eaves, 1977) were extremely influential in clarifying the properties of existing analytic approaches to family studies that use raw data, correlations or mean squares as the starting point. They also introduced the use of covariance matrices as an alternative and integrated factor analysis methodology into biometrical genetic analysis. These two papers, together with Martin et al. (1978), laid much of the foundation for the later developments in human behavior genetics, including the establishment of large twin registries and the development of modern maximum likelihood approaches for model estimation and testing that enabled the extension of the classical twin model to threshold traits, multiple phenotypes and extended twin-families (Neale & Cardon, 1992).

Power calculation has remained an important issue in human genetics research. Subsequent papers to Martin et al. (1978) have considered the power of new study designs including threshold traits (Neale et al., 1994), multivariate phenotypes (Schmitz et al., 1998) and extended twin designs (Posthuma & Boomsma, 2000). As the field moved to include molecular data for gene mapping, analytic power calculations were developed for quantitative trait linkage and association analyses under the variance components model, also using the noncentral chi-squared distribution (Nance & Neale, 1989; Purcell et al., 2003; Sham et al., 2000). In the genome-wide association studies (GWAS) era, the variance components model has been applied to estimate the heritability attributable to common single-nucleotide polymorphisms (SNPs), and the power of this approach has also been characterized (Visscher et al., 2014).

Where the noncentral, chi-squared distribution is a poor approximation of the sampling distribution of the test, simulation-based approaches to power calculation can be used. Of course, all power calculations are effectively simulations, where expected values of statistics such as mean squares or covariances are generated from known values of the parameters of the model in question. Fitting models to summary statistics in this way is very efficient because only two models need to be fitted to the data — the true one and a submodel where one or more of the parameters have been fixed to zero. An extension of this method is to generate raw data and to fit the true and the false models to them. This approach is more flexible because it allows datasets with many patterns of missing observations to be handled with ease. Similarly, models with definition variables can be tested with this approach. A key consideration here is whether to generate data that exactly conform to the covariance matrix and means used to simulate them (e.g., using argument empirical = TRUE in the mvrnorm() routine in the MASS R library; Venables & Ripley, 2002). Doing so reduces the computational complexity to simulating and model-fitting to only one raw dataset. At other times, permitting stochastic variation in the generation of datasets can be useful, particularly when the statistics used to evaluate model fit do not conform to, for example, the noncentral chi-squared distribution. The multivariate ACE Cholesky is one example. Here, large numbers of trials of simulating and fitting are needed to characterize the distribution of the trials' test statistics. Having done so, it is then possible to evaluate the probability of observing effect sizes that exceed a particular threshold on the empirical distribution of the test statistics. This procedure aligns closely with using the bootstrap likelihood ratio test (BLRT; Boker et al., 2020).

The seminal paper of Martin et al. (1978) on the power of the classical twin design was revisited by Visscher (2004), who calculated power via the standard errors of the variance components and the expected values of the maximum likelihood ratio test statistics. His results are largely comparable to those of Martin et al. (1978), with the major difference being that the consideration of likelihood ratio statistics enabled a specific parameter in a model to be tested (e.g., the additive genetic effects within a full model that also contains shared sibship environment and individual-specific environment), rather than the entire model. It should be noted that such calculations are not limited to estimating the power to detect nonzero heritability. Estimates of the power to detect other variance components, particularly that due to the shared or family environment (Visscher et al., 2008), are also of great utility.

An important further consideration in statistical power is whether variance components should be bounded to be greater than zero. If a model's variance components are estimated directly and without bounds, they may return nonsensical negative values. However, this difficulty in interpretability is counterbalanced by the good statistical properties of the noncentrality parameter, which can be interpreted without transformation. Of note, in recent work by Verhulst et al. (2019), is that models that do bounds can deliver very poor estimates of statistical power, unless challenging mathematical transformations are carried out (Wu & Neale, 2013). Some prior discrepancies between the articles emerge as a result of model constraints; for example, estimating the path coefficient from genotype to phenotype versus estimating the variance component VA and allowing this quantity to be negative. This issue becomes much more serious in power calculations for multivariate genetic models. That Nick's 1978 paper and thesis have led to new studies on the topic over 40 years later is a great tribute to his ability to produce useful science that stands the test of time.

By highlighting statistical power considerations, Nick calls to mind Ronald Fisher (1938), who in his Presidential Address to the First Indian Statistical Congress said 'to consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the

https://doi.org/10.1017/thg.2020.46

experiment died of' (p. 17). The power calculations of Martin and colleagues are exactly the kind of prospective treatments that can prevent horribly underpowered research studies from being carried out. While meta-analysis can overcome some shortcomings of studies that involve too few subjects, most researchers would prefer to have results from adequately powered studies that can contribute substantively either alone or in aggregate with others. Power calculations can take much of the guesswork out of research planning. In some cases, logistical considerations place additional constraints on the maximum sample size that can practically be collected. Power calculations remain useful here — at the very least to avoid proceeding with a study where all the findings will be equivocal and difficult to validate. The International Methodology Workshops taught in Europe and Colorado continue to teach methodology for statistical power calculations for exactly this reason, which is but one reflection of an enduring contribution by Dr Martin.

As a pioneer of the fields of biometrical and behavioral genetics, Nick's knowledge, insights and perspectives have benefitted entire generations of researchers in behavioral genetics who have attended the annual 'twin workshops,' often multiple times. We were fortunate to progress to faculty members of the workshop and have more directly experienced Nick's enthusiasm and intellectual curiosity, greatly facilitating the sharing of ideas and lively debates, not only among faculty members but also with the students. These debates and discussions were what have made the workshops so enjoyable and often led to new and fruitful research directions. On the occasion of Nick's 70th birthday, we express our appreciation and gratitude to him, glance backward to what we have achieved and look forward to working together to extend the frontiers of the field.

References

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., … Niesen, J. (2020). *OpenMx 2.16.0 User Guide*. Open MX. https://openmx.ssri.psu.edu/.

Eaves, L. J., Last, K. A., Young, P. A., & Martin, N. G. (1978). Model-fitting approaches to the analysis of human behaviour. *Heredity*, *41*, 249–320.

Evans, D. M., Gillespie, N. A., & Martin, N. G. (2002). Biometrical genetics. *Biological Psychology*, *61*, 33–51.

Fisher, R. A. (1938). Presidential address to the first Indian statistical congress. *Sankhyā: The Indian Journal of Statistics (1933–1960)*, *4*, 1–17.

Jinks, J. L., & Fulker, D. W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin*, *73*, 311–349.

Martin, N. G. (1976). The classical twin study in human behavioural genetics (Unpublished doctoral thesis). University of Birmingham.

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity*, *38*, 79–95.

Martin, N. G., Eaves, L. J., Kearsey, M. J., & Davies, P. (1978). The power of the classical twin study. *Heredity*, *40*, 97–116.

Mather, K., & Jinks, J. L. (1982). *Biometrical genetics* (3rd ed.). London, UK: Chapman and Hall.

Nance, W. E., & Neale, M. C. (1989). Partitioned twin analysis: A power study. *Behavior Genetics*, *19*, 143–150.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families*. Dordrecht, Germany: Kluwer Academic Press.

Neale, M. C., Eaves, L. J., & Kendler, K. S. (1994). The power of the classical twin study to resolve variation in threshold traits. *Behavior Genetics*, *24*, 239–258.

Posthuma, D., & Boomsma, D. I. (2000). A note on the statistical power in extended twin designs. *Behavior Genetics*, *30*, 147–158.

Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics (Oxford, England)*, *19*, 149–150.

Schmitz, S., Cherny, S. S., & Fulker, D. W. (1998). Increase in power through multivariate analyses. *Behavior Genetics*, *28*, 357–363.

Sham, P. C., Cherny, S. S., Purcell, S., & Hewitt, J. K. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics*, *66*, 1616–1630.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer. http://www.stats.ox.ac.uk/pub/MASS4

Verhulst, B., Prom-Wormley, E., Keller, M., Medland, S., & Neale, M. C. (2019). Type I error rates and parameter bias in multivariate behavioral genetic models. *Behavior Genetics*, *49*, 99–111.

Visscher, P. M. (2004). Power of the classical twin design revisited. *Twin Research*, *7*, 505–512.

Visscher, P. M., Gordon, S., & Neale, M. C. (2008). Power of the classical twin design revisited: II detection of common environmental variance. *Twin Research and Human Genetics*, *11*, 48–54.

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A., Chen, G. B., Lee, S. H., Wray, N. R., Goddard, M. E., & Yang, J. (2014). Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genetics*, *10*, e1004269.

Wu, H., & Neale, M. C. (2013). On the likelihood ratio tests in bivariate ACDE models. *Psychometrika*, *78*, 441–463.

https://doi.org/10.1017/thg.2020.46