A survey of current software for genetic power calculations

Jo Knight

Social Genetic and Developmental Psychiatry Research Centre, Box P080, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London, SE5 8AF, UK;
Tel: +44 (0)20 7848 0854; Fax: +44 (0)20 7848 0866; E-mail: J.Knight@iop.kcl.ac.uk

Date received (in revised form): 2nd March, 2004

Abstract

Estimation of power is a key step in any study. This review briefly outlines the factors that affect power and the two main approaches for estimating it. There are a number of web-based tools and programs freely available to enable geneticists to perform power calculations, and the specifics of some of these are discussed here.

Keywords: *power, software*

Introduction

The power of a study is the probability that it will detect an effect of a given size, and is therefore a subject of great importance. It is related to the magnitude of the effect, the sample size and the chosen level of statistical significance (ie the probability of a false-positive result). Ideally, calculations are carried out in the early stages of planning, in order to establish the number of people required.

In genetic studies, power is estimated either by asymptotic approaches or by undertaking simulations. The former involves employing closed equations, whereas the latter requires the creation of thousands of datasets with the same parameters as the population being studied. (The proportion of simulated sets yielding positive analysis results gives an estimate of the power.) Simulation can be a more accurate approach than the use of closed equations if the investigator is able to use the correct parameters. As the parameters required (eg the frequency of the causative variant) are often unknown, however, this is by no means an inconsequential task. Furthermore, simulation approaches are usually more computer intensive and time consuming. Both approaches are required because of the diversity of calculations performed in the context of genetic studies. Where asymptotic methods have not been established, or for some reason are not considered sufficient, simulation can be used.

Despite the complexities, a variety of tools have been designed which allow investigators to estimate power using closed equations and/or to simulate a wide range of datasets. The purpose of this paper is to outline a number of these freely available programs and web-based utilities. Box 1 provides a summary of the tools, highlighting the nature of each utility, where they can be downloaded from and brief information about what they can do.

The range of types of software available is the first thing to note. As well as stand-alone programs, there are web-based tools and downloadable Excel spreadsheets. Some are designed to perform simulations and some to calculate power from closed equations, others perform both tasks in addition to data analysis.

The software

SIMLINK and SLINK, written in 1990 and 1991, respectively, are the tools that have been available for the longest time.[1–4] Both are stand-alone programs and allow the user to carry out simulation studies on pedigrees to establish power for parametric linkage analysis; hence, they require the same information about the trait under study that is requisite to such analysis. Since the development of these tools, a number of other simulation programs, with different requirements, have been written; for example, ASP, SIMLA, SIMNUC and GASP. Such programs can be used to assess the power of nonparametric linkage studies. In addition, the closed equations derived in 1990 by Risch[5] to calculate power for studies of affected siblings are programmed into a spreadsheet called POWTEST, available from Dave Curtis's website.

Closed equations for the detection of both linkage and association using variance components analysis have been encoded in the Genetic Power Calculator (GPC).[6] Furthermore, GPC has an option to estimate the contribution to the

Box 1. Summary of available tools

ASP, SIMLA, SIMLINK, SIMNUC, SLINK

- Downloadable programs
- Simulation of pedigrees
- Respective website addresses:
    - http://www.uni-kiel.de/medinfo/mitarbeiter/krawczak/download/index.html
    - http://www.chg.duke.edu/software/simla.html
    - http://www.sph.umich.edu/csg/boehnke/simlink.html
    - http://linkage.rockefeller.edu/ott/simnuc.html
    - http://watson.hgen.pitt.edu/fom-serve/cache/20.html

Genetic Power Calculator (GPC)

- Web-based utility
- Closed equations for linkage and association of qualitative or quantitative traits in the variance components framework; power for individual sibships with trait data and case-control; and TDT for binary traits and threshold-selected traits
- http://statgen.iop.kcl.ac.uk/gpc/

Merlin-Regress

- Downloadable program
- Closed equations for expected LOD scores based on regression approaches
- http://www.sph.umich.edu/csg/abecasis/Merlin/

Power for Association With Errors (PAWE)

- Web-based utility
- Closed equations for case-control association with errors
- http://linkage.rockefeller.edu/joanne/pawe/

PBAT

- Downloadable program
- Closed equation, simulation and analysis for family-based association studies
- http://www.biostat.harvard.edu/~clange/default.htm

POWER

- Downloadable program
- Closed equations for studies of interactions
- http://dceg2.cancer.gov/POWER/

POWTEST

- Excel spreadsheet
- Closed equations for TDT and linkage with affected sib pairs
- http://www.mds.qmw.ac.uk/statgen/dcurtis/software.html

QUANTO

- Downloadable program
- Closed equations for studies of interactions
- http://hydra.usc.edu/gxe/

TDT calculator

- Downloadable program
- Closed equation and simulation for family-based association studies
- http://biosun01.biostat.jhsph.edu/~wmchen/pc.html

UCLA stat calculator

- Web-based utility
- Closed equation for case control association (as well as closed equations for other non-genetic study types)
- http://calculators.stat.ucla.edu/powercalc/binomial/case-control/b-case-control-power.php

TDT, transmission disequilbrium test

test statistic of each sibship using trait data.[7] This allows ranking of sibships and hence provides a way of prioritising genotyping. An extension of this method is implemented in Merlin-Regress, where the expected LOD scores can be calculated for general pedigrees.[8] Merlin-Regress is also able to perform regression-based analysis for quantitative traits in phenotypically selected samples.

The GPC is perhaps the utility capable of performing the widest range of power calculations. In addition to the utilities already mentioned, it can also be used to calculate power for transmission disequilibrium tests (TDTs) of binary traits and TDT and case-control studies of threshold-selected quantitative traits. Calculating power for these tests in the GPC is advantageous, as the GPC takes linkage disequilibrium between the gene and the marker under study into account. This web-based utility calculates power from the information provided by the user and produces output that is concise and useful. Accompanying notes relate mainly to usage rather than theory, and direct the user to papers in which the latter is explained.

Family-based association studies are frequently used for gene mapping. Extensions of TDT allow for analysis of quantitative as well as dichotomous traits; inclusion of families with missing parents; and joint analysis of different types of families (eg single affected/multiple affected and discordant siblings). PBAT[9,10] and the TDT[11] calculator allow the user to perform closed-form calculations and simulation for such studies. The closed equations are slightly different. In the paper that outlines the theory behind PBAT, the authors suggest their approach is more accurate than that of Chen, as it calculates the power of the actual test statistic whereas Chen computes the power of the expected statistic.[10] Lange and Laird suggest that, although this does not appear to make a lot of difference in smaller studies, there is a greater difference in large studies.[10]

Both PBAT and TDT are stand-alone programs. PBAT has a very helpful and detailed web page that includes everything from downloading instructions to an explanation of how to use the program. Furthermore, PBAT can actually carry out family-based association tests. There is no documentation for the TDT calculator but it is easy to use.

Researchers are becoming increasingly interested in investigating the combined effects of genetics and the environment, as well as the interactions between different genes. At least two programs are available to calculate power for such studies, Quanto[12,13] and a National Cancer Institute program called 'Power'.[14] These programs are designed for regression-based approaches. Quanto has the advantage of dealing with a wider range of study designs, including certain family-based populations as well as quantitative traits.

The final program that will be introduced here is Power Association With Errors (PAWE).[15,16] This web-based utility, available on the Rockefeller website, incorporates an error model into its power calculations. It computes power and sample size calculations for genetic case-control association studies in the presence of genotyping errors, and determines how much genotyping errors cost the researcher, in terms of decreased asymptotic power for a fixed sample size or increased sample size, to maintain constant asymptotic power.

This paper covers a variety of useful tools which should be helpful to geneticists attempting to perform power calculations; however, it is important not to become complacent. These calculations are, at best, an estimate of the power of the study, as the parameters used in them are often unknown. Furthermore, they will be imprecise when they do not take into account all of the factors that influence the magnitude of the effect. It is, therefore, encouraging to find recent programs, like PAWE, which continue to take steps to improve accuracy.

References

1. Boehnke, M. (1986) 'Estimating the power of a proposed linkage study: A practical computer simulation approach', *Am. J. Hum. Genet.* Vol. 39, pp. 513–527.
2. Ploughman, L.M. and Boehnke, M. (1989) 'Estimating the power of a proposed linkage study for a complex genetic trait', *Am. J. Hum. Genet.* Vol. 44, pp. 543–551.
3. Ott, J. (1989) 'Computer-simulation methods in human linkage analysis', *Proc. Natl Acad. Sci. USA* Vol. 86, pp. 4175–4178.
4. Weeks, D.E., Ott, J. and Lathrop, G.M. (1990) 'SLINK: A general simulation program for linkage analysis', *Am. J. Hum. Genet.* Vol. 47, p. A204.
5. Risch, N. (1990) 'Linkage strategies for genetically complex traits. II. The power of affected relative pairs', *Am. J. Hum. Genet.* Vol. 46, pp. 229–241.
6. Purcell, S., Cherny, S.S. and Sham, P.C. (2003) 'Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits', *Bioinformatics* Vol. 19, pp. 149–150.
7. Purcell, S., Cherny, S.S., Hewitt, J.K. and Sham, P.C. (2001) 'Optimal sibship selection for genotyping in quantitative trait locus linkage analysis', *Hum. Hered.* Vol. 52, pp. 1–13.
8. Sham, P.C., Purcell, S., Cherny, S.S. and Abecasis, G.R. (2002) 'Powerful regression-based quantitative-trait linkage analysis of general pedigrees', *Am. J. Hum. Genet.* Vol. 71, pp. 238–252.
9. Lange, C., DeMeo, D.L. and Laird, N.M. (2002) 'Power and design considerations for a general class of family-based association tests: Quantitative traits', *Am. J. Hum. Genet.* Vol. 71, pp. 1330–1341.
10. Lange, C. and Laird, N.M. (2002) 'Power calculations for a general class of family-based association tests: Dichotomous traits', *Am. J. Hum. Genet.* Vol. 71, pp. 575–584.
11. Chen, W.M. and Deng, H.W. (2001) 'A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes', *Genet. Epidemiol.* Vol. 21, pp. 53–67.
12. Gauderman, W.J. (2002) 'Sample size requirements for matched case-control studies of gene–environment interaction', *Stat. Med.* Vol. 21, pp. 35–50.
13. Gauderman, W.J. (2002) 'Sample size requirements for association studies of gene–gene interaction', *Am. J. Epidemiol.* Vol. 155, pp. 478–484.
14. Garcia-Closas, M. and Lubin, J.H. (1999) 'Power and sample size calculations in case-control studies of gene–environmental interactions: Comments on different approaches', *Am. J. Epidemiol.* Vol. 148, pp. 689–693.
15. Gordon, D., Finch, S.J., Nothnagel, M. and Ott, J. (2002) 'Power and sample size calculations for case-control genetic association tests when errors present: Application to single nucleotide polymorphisms', *Hum. Hered.* Vol. 54, pp. 22–33.
16. Gordon, D., Levenstien, M.A., Finch, S.J. and Ott, J. (2003) 'Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies', *Pac. Symp. Biocomput.* Vol. 8, pp. 490–501 at http://psb.stanford.edu/psb-online