StatNews #71
Sample Size Calculation in Genetics Studies
December 2007

Note: As of 2012 Quanto is no longer the most current tool for sample size calculation in genetic association studies. More current open-source options for computing sample size for genetic studies are available within a host of R packages. The most comprehensive repository for searching out the most up to date R packages in this area is http://www.bioconductor.org/.

For common complex chronic diseases, genetic association studies have the potential to assess the associations between disease status and genetic variants in a population. As with other studies calculating the appropriate sample size is an important part of the study design (see: http://www.cscu.cornell.edu/news/statnews/stnews41.pdf). The calculation of the sample size for a genetic study requires however some additional considerations since one needs to take into account the inheritance model (dominant, recessive or additive), the frequency of the high risk alleles in the population, the overall disease risk in the population, in addition to the choice of an odds ratio (or relative risk), the power and the significance level. In this newsletter we introduce a Window-based software package, Quanto, developed by Jim Gauderman, PhD, and John Morrison, M.S. of University of Southern California, which can be used to compute sample size and power in genetics studies.

Although there are a number of software packages available in genetic studies for sample size or power calculation, "Quanto" is to be recommended because of its efficiency and user-friendliness. It is a menu driven software package which is available free of charge and can be downloaded from http://hydra.usc.edu/gxe/. Two types of outcome can be considered: a disease (binary) outcome and a quantitative (continuous) outcome.

Available study designs include the case-control, case-sibling, case-parent, and case-only designs. Case-control studies use patients who have a disease and look back to see if the characteristics of the affected patients are different from those who don't have the disease. Suppose, in a case control study, DNA samples have been collected to determine the effects of each SNP's1 on the risk of having cardiovascular disease. We are interested in calculating the sample size needed to have the effect size (or odds ratio) in the range of 1.5-2.0 with at least 80 percent power under a dominance model. Moreover, the minor allele frequency2 is chosen to be 10 percent, and a type 1 error level of 0.05. Once these values are entered, Quanto can compute the required sample size.

The following link has an example to see how Quanto works.
http://www.cscu.cornell.edu/news/statnews/stnews71Quantoexample.pdf
Quanto can also handle more complex genetics association studies such as those involving a gene-gene interaction in which different genes are involved in common pathways to disease. In addition,

for some diseases there may exist gene-environment interactions For example, the genes may be expressed differently depending on the diet of individuals.

The following references are useful if you would like to learn more about computing sample size and power in genetics studies. Please do not hesitate to contact the Cornell Statistical Consulting Unit if you would like assistance with the sample size calculation of a genetic study.

References:

1. WJ Gauderman (2002), "Sample size requirements for matched case-control studies of gene-environment interaction", Stat Med 21:35-50.
2. WJ Gauderman (2002), "Sample size requirements for association studies of gene-gene interaction", American Journal of Epidemiology, 155:478-484.
3. WJ Gauderman (2003), "Candidate gene association studies for a quantitative trait, using parent-offspring trios", Genetic Epidemiology, 25:327-338.

Author: Resmi Gupta

[1] DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. (Source: www.biochem.northwestern.edu)

[2] A measure of how common an allele is in a population

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).